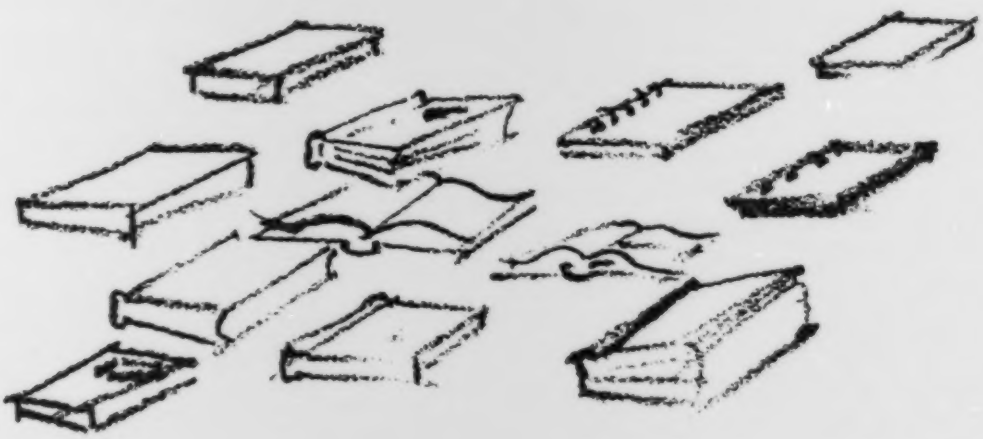
Statistiques

DOCUMENTS DE RÉFÉRENCE

RECUEIL DES TEXTES DES PRÉSENTATIONS DU COLLOQUE SUR LES MÉTHODES ET DOMAINES D'APPLICATION DE LA STATISTIQUE

1989



Bureau de
la statistique
du Québec

90-04761/

Pour tout renseignement concernant le BSQ
et les données statistiques qui y sont disponibles,
s'adresser au:

Bureau de la statistique du Québec
117, rue Saint-André
Québec (Québec)
G1K 3Y3
Tél.: (418) 691-2401

ou

Tél.: 1-800-463-4090
(aucuns frais d'appel)

Dépôt légal
Bibliothèque nationale du Québec
Deuxième trimestre 1990
ISBN 2-551-12348-8

© Bureau de la statistique du Québec

Toute reproduction est interdite
sans l'autorisation expresse
du Bureau de la statistique du Québec.

Avant-propos

Ce recueil contient les textes des présentations du colloque "Méthodes et domaines d'application de la statistique 1989" qui s'est tenu à l'Université du Québec à Montréal en mai 1989, dans le cadre du 57e congrès de l'Association canadienne-française pour l'avancement des sciences. Pour une quatrième année consécutive, le colloque était parrainé par le Bureau de la statistique du Québec. La diversité des organismes ayant participé au colloque et des thèmes présentés lui assurèrent un franc succès. En effet, 14 séances en statistique regroupant plus de 70 conférenciers ont été suivies par au-delà de 160 participants de tous les milieux de la statistique.

Nous tenons à remercier tous ceux qui nous ont fait parvenir le texte de leur présentation. Nous espérons que la diffusion de ce recueil permettra aux statisticiens de profiter des expériences et recherches ici démontrées. Le Bureau de la statistique du Québec laisse aux auteurs la responsabilité entière du contenu de leur texte.

Le directeur général

Luc Bessette

Responsable du colloque et
coordonnatrice de cette publication: Françoise Tarte
Statisticienne
Direction de la méthodologie

Bureau de la statistique du Québec
117, rue St-André
Québec (Québec)
Tél.: (418) 691-2411

Table des matières

OUVERTURE DU COLLOQUE

BESSETTE, Luc, Directeur général, Bureau de la statistique du Québec. Discours d'ouverture..... 3

SESSIONS SUR LA MÉTHODOLOGIE D'ENQUÊTES

BÉLAND, Yves, Statistique Canada.
Utilisation des tests aléatoires dans le cadre de la planification du recensement de 1991..... 15

DAOUST, Pierre, Statistique Canada.
La pondération lors du recensement de la population de 1986..... 23

DUMAIS, Jean, HIDIROGLOU, Michel A., Statistique Canada.
Le système général d'estimation pour les enquêtes économiques infra-annuelles..... 45

GWET, Jean-Philippe, Université Laval.
De la robustesse dans l'estimation de la moyenne d'une population finie: une généralisation de la méthode du quotient..... 57

HIDIROGLOU, Michel A., Statistique Canada.
PC CARP: Un logiciel pour le calcul de la variance dans les enquêtes complexes..... 69

JONCAS, Marc, Statistique Canada.
Évaluation de la méthode de sélection de l'échantillon au recensement de la population canadienne..... 79

JULIEN, Claude, MARANDA, François, Statistique Canada.
Le remaniement du plan de sondage de l'enquête nationale sur les fermes de 1988..... 89

MASSE, Sylvain, SOM.
Estimation de la moyenne et de la variance d'échantillonnage pour un plan quelconque à l'aide du Jackknife..... 111

TAMBAY, Jean-Louis, Statistique Canada.
Une approche intégrée du traitement des valeurs aberrantes dans les enquêtes économiques infra-annuelles..... 133

VÉGIARD, Sylvain, Régie des assurances agricoles du Québec.
Problèmes d'échantillonnage rencontrés dans l'administration de programmes d'assurance-récolte..... 145

SESSIONS SUR L'ANALYSE DE DONNÉES

BILOCQ, France, BERTHELOT, Jean-Marie, Statistique Canada.
Un schéma de vérification basé sur l'analyse multivariée.. 157

CLOUTIER, Esther, LAFLAMME, Lucie, Institut de recherche en santé et en sécurité du travail du Québec.
Diagnostic quantitatif et qualitatif des zones de risques en scieries..... 167

DAIGLE, Gaétan, Université Laval.
Représentation biplot robuste..... 177

DUFOUR, Sandra, D'AVIGNON, Gilles, Université Laval.
Mesures de similitude dans l'analyse de regroupement d'hôpitaux..... 187

LAPOINTE, France, Université Laval.
Effet de l'utilisation d'une échelle de réponses dichotomique dans la sélection d'un vecteur de poids..... 201

LAPOINTE, François-Joseph, Université de Montréal.
La comparaison statistique de classifications hiérarchiques..... 213

LAVIGNE, Mylène, Statistique Canada.
Modèles log-linéaires pour des tableaux de contingence complets..... 223

PLANTE, Nathalie, RIVEST, Louis-Paul, Université Laval.
Estimation de la taille d'une population de saumonneaux à l'aide d'un modèle de capture-recapture avec stratification..... 239

SÄRNDAL, Carl-Erik, Université de Montréal, SWENSSON, Bengt, Université d'Örebro et WRETMAN, Jan H., Statistique Suède.
Estimation de la variance de l'estimateur par régression du total d'une population finie par la méthode des résidus pondérés..... 251

SESSION SUR LES APPLICATIONS DE LA STATISTIQUE DANS LE DOMAINE DES TRANSPORTS

BERGERON, Paul, ministère des Transports du Québec.
Évaluation de l'état du réseau routier à l'entretien du ministère des Transports..... 277

DÉSILETS, Martine, OUIMET, Claude, ministère des Transports du Québec.
Identification de profils types de circulation pour fins d'estimation..... 285

SESSION SUR LA CONFIDENTIALITÉ DES DONNÉES ET LE COUPLAGE DE FICHIERS ADMINISTRATIFS

DENIS, Johanne, Statistique Canada.
Fichiers publics de microdonnées: le problème de l'identification et de la divulgation et quelques techniques palliatives..... 297

SESSION SUR LES SÉRIES CHRONOLOGIQUES

BÉGIN, Ginette, LATOUR, Alain, Université du Québec à Montréal, CHANUT, Jean-Pierre, EL-SABH, M.I., Université du Québec à Rimouski.
Modélisation multivariée du courant de Gaspé..... 307

BRABANT, Carole, LATOUR, Alain, FERRARIS, Jocelyne, MERGLER, Donna, Université du Québec à Montréal.

Étude comparative de modèles chronologiques décrivant la réponse cardiaque en situation réelle de travail..... 315

LABERGE, Claude, Institut national de recherche scientifique-Eau.

Application d'estimateurs robustes pour le traitement de séries chronologiques environnementales..... 329

LATOUR, Alain, Université du Québec à Montréal.

Le choix d'une méthode de prévision..... 339

LEFRANÇOIS, Bernard, Statistique Canada.

La détection des observations surinfluentes dans les séries chronologiques..... 345

POIRIER, Claude, LANIEL, Normand, Statistique Canada.

L'ajustement des données annuelles de flux à l'année civile..... 357

QUENNEVILLE, Benoît, DAGUM, Estela Bee, Statistique Canada.

Problèmes pratiques associés au calcul de la variance des données désaisonnalisées par X-11-Arima..... 367

RAYNAULD, Jacques, SIMONATO, Jean-Guy, École des Hautes études Commerciales.

Estimation de modèles avec variables non observables dynamiques: quelques illustrations utilisant la méthode EM..... 377

SESSION SUR LES PLANS D'EXPERIENCE

LALONDE, Yves, Laboratoires Nordic.
Méthodes pour traiter l'interaction dans un plan d'étude d'un essai clinique..... 389

MONTÉGIANI, Mario, ministère des Transports du Québec.
Détermination des causes d'orniérage à l'aide de la planification d'expérience (suite et fin)..... 411

SESSION SUR LA BIOSTATISTIQUE

BRAVO, Gina, Hôpital d'Youville de Sherbrooke.
Estimation simultanée de la fiabilité intra-juges et inter-juges..... 425

CARDIN, Sylvie, SUISSA, Samy, BOIVIN, Jean-François, Université McGill et Département de santé communautaire de l'hôpital général de Montréal.
Estimation pour petites régions en santé communautaire.... 439

CIAMPI, Antonio, THIFFAULT, Johanne, du BERGER, Roxane, Université McGill.
SEGMAG: Analyse de données par construction d'arbres en biostatistique..... 447

LAMBERT, Jean, Université de Montréal.

Stratégies d'inférence statistique en recherche épidémiologique en présence de paramètres de nuisance..... 469

SESSION SUR LA STATISTIQUE NON PARAMÉTRIQUE

PERREAULT, Luc, Université Laval.

Estimation et intervalle de confiance adaptable pour un paramètre de localisation..... 487

Ouverture du colloque

DISCOURS D'OUVERTURE
COLLOQUE MÉTHODES ET DOMAINES D'APPLICATION DE LA STATISTIQUE 1989 - 57E CONGRÈS DE L'ACFAS

LUC BESSETTE
DIRECTEUR GÉNÉRAL
BUREAU DE LA STATISTIQUE DU QUEBEC

MESDAMES ET MESSIEURS,

PERMETTEZ-MOI À MON TOUR DE VOUS SOUHAITER LA BIENVENUE À CE 57E CONGRÈS DE L'ACFAS ET PLUS SPÉCIFIQUEMENT AU COLLOQUE SUR LES MÉTHODES ET DOMAINES D'APPLICATION DE LA STATISTIQUE 1989. GRÂCE À VOTRE PARTICIPATION, CET ÉVÉNEMENT EST DEVENU AVEC LES ANNÉES UN MOMENT PRIVILÉGIÉ POUR LES MÉTHODOLOGISTES ET LES STATISTICIENS DE LANGUE FRANÇAISE. VOILÀ POURQUOI LE BUREAU DE LA STATISTIQUE DU QUÉBEC EST FIER DE S'ASSOCIER À L'ACFAS EN PARRAINANT POUR LA QUATRIÈME ANNÉE CONSÉCUTIVE CE COLLOQUE. ÉTANT DONNÉ QUE VOTRE INTÉRÊT EST SURTOUT POUR LES PRÉSENTATIONS DES DIFFÉRENTS CONFÉRENCIERS ET NON POUR LE DISCOURS D'OUVERTURE, JE TENTERAI D'ÊTRE LE PLUS BREF POSSIBLE.

JE VOUDRAIS, DANS UN PREMIER TEMPS, REMERCIER ET SOULIGNER LE TRAVAIL DE LA RESPONSABLE DE L'ORGANISATION DE COLLOQUE, MADAME FRANÇOISE TARTE. C'EST PLUSIEURS HEURES DE TRAVAIL QUE FRANÇOISE A CONSACRÉES À CE PROJET. ELLE MÉRITE TOUTE NOTRE RECONNAISSANCE. VOUS SAVEZ, ON PREND TROP SOUVENT POUR ACQUIS LORSQUE L'ON PARTICIPE À UN ÉVÉNEMENT COMME CELUI-CI LE TRAVAIL D'ORGANISATION. TOUT SEMBLE SI NATUREL LORSQUE L'ON

S'Y PRÉSENTE. CROYEZ-MOI, C'EST UN TRAVAIL ARDU QUI DEMANDE BEAUCOUP DE DISPONIBILITÉ ET UN SENS AIGUISÉ DE L'ORGANISATION. VOUS SAVEZ, LA DIFFÉRENCE ENTRE UNE BONNE IDÉE ET UN ÉVÉNEMENT RÉUSSI EST SOUVENT TOUT CE TRAVAIL QUI SE FAIT EN COULISSE.

MALHEUREUSEMENT POUR LES ORGANISATEURS, LORSQU'UN COLLOQUE EST RÉUSSI ON A TENDANCE À EN DONNER TOUT LE MÉRITE AUX CONFÉRENCIERS, MAIS LORSQU'IL Y A DES RATÉS ON BLÂME VITE L'ORGANISATION. EN RÉALITÉ LE SUCCÈS D'UN COLLOQUE TIENT À LA QUALITÉ DES CONFÉRENCIERS ET DES EXPOSÉS, À LA PARTICIPATION DES CONGRESSISTES ET À L'EFFICACITÉ ET LA CRÉATIVITÉ DES ORGANISATEURS.

VOILÀ POURQUOI JE VOUDRAIS AUSSI REMERCIER LES CONFÉRENCIERS QUI ONT ACCEPTÉ L'INVITATION DE LA RESPONSABLE DU COLLOQUE. VOTRE PARTICIPATION EST ESSENTIELLE. VOS EXPOSÉS ALIMENTERONT AU COURS DES PROCHAINS JOURS LES ÉCHANGES ET LES DISCUSSIONS DE CE COLLOQUE. LA MEILLEURE ORGANISATION AU MONDE NE POURRAIT ASPIRER AU SUCCÈS SANS LA PARTICIPATION DE CONFÉRENCIERS DE QUALITÉ ET LA PRÉSENTATION D'EXPOSÉS DE QUALITÉ. UNE LECTURE RAPIDE DU PROGRAMME NOUS MONTRE QUE NOUS SERONS GÂTÉS AU COURS DES PROCHAINS JOURS.

FINALEMENT, JE VEUX REMERCIER TOUS LES PARTICIPANTS. VOTRE PRÉSENCE EST LE PRÉTEXTE À LA TENUE D'UN TEL COLLOQUE. DU DEGRÉ ET DE LA QUALITÉ DE VOTRE PARTICIPATION DÉPENDENT LE DYNAMISME ET L'UTILITÉ DE CE COLLOQUE.

Ce colloque se veut un lieu d'échange privilégié sur les derniers travaux et développements dans le domaine de la méthodologie statistique. Il est l'occasion pour les statisticiens et méthodologistes d'expression française d'exposer l'objet de leur recherche des dernières années, les développements ou applications qui en ont suivis et les analyses ou orientations qui en découlent. Ce colloque permet surtout aux statisticiens et méthodologistes des administrations publiques, des universités et du secteur privé d'échanger et de partager leurs expériences. Il y a d'ailleurs ici un représentant de l'INED, Monsieur Rienday, qui est ici pour étudier la formule de ce colloque.

C'est plus de 70 exposés qui vous seront présentés au cours de ces quatre jours dans des domaines aussi variés que la méthodologie d'enquêtes, les méthodes non paramétriques, l'analyse de données, la biostatistique, les séries chronologiques, le contrôle statistique de la qualité, la confidentialité et le couplage de fichiers administratifs, les systèmes généraux de collecte, d'imputation et de traitement des données, les plans d'expérience et l'utilisation des statistiques.

C'est plus de 95 conférenciers qui vous entretiendront au cours des prochains jours. Ils nous viennent de la plupart des universités québécoises ou de leurs constituantes, de plusieurs ministères et organismes des secteurs publics, parapublics et puripublics québécois et du secteur privé.

MAIS POURQUOI DONNER TANT D'IMPORTANCE À LA STATISTIQUE ET PLUS SPÉCIFIQUEMENT À LA MÉTHODOLOGIE ET AU DOMAINE D'APPLICATION DE LA STATISTIQUE? POURQUOI UN TEL COLLOQUE SUSCITE-T-IL UN TEL NIVEAU DE PARTICIPATION TANT AU NIVEAU DES CONFÉRENCIERS QUE DES INSCRIPTIONS? POURQUOI UN COLLOQUE POUR LES MÉTHODOLOGISTES ET STATISTICIENS DE LANGUE FRANÇAISE?

IMPORTANCE DES MÉTHODES ET DOMAINES D'APPLICATION DE LA STATISTIQUE

PERMETTEZ AU GESTIONNAIRE QUE JE SUIS DE TENTER DE RÉPONDRE À CETTE QUESTION. IL Y A D'ABORD BIEN SÛR LES PROGRÈS ACCOMPLIS DANS CE CHAMP DE LA SCIENCE. UN PEU COMME L'INFORMATIQUE, LA MÉTHODOLOGIE STATISTIQUE EST UN DOMAINE EN PLEINE ÉVOLUTION, UN OUTIL À CARACTÈRE UN PEU MAGIQUE ET LA PLANCHE DE SALUT DE LA STATISTIQUE FACE AUX COMPRESSIONS BUDGÉTAIRES, TOUT EN DEVENANT UN BIEN DE CONSOMMATION JOURNALIÈRE QUI GAGNE CHAQUE JOUR UN PEU PLUS DE CRÉDIBILITÉ.

LES EXEMPLES LES PLUS ÉVIDENTS SONT SANS AUCUN DOUTE LES SONDAGES D'OPINION. D'OPÉRATIONS MYSTÉRIEUSES FRÔLANT LA SORCELLERIE ET SUSCITANT PEU DE CRÉDIBILITÉ, ILS SONT DEVENUS AU COURS DES DERNIÈRES ANNÉES DES OUTILS DE TRAVAIL COURUS NON SEULEMENT PAR LES POLITICIENS ET LES PARTIES POLITIQUES, MAIS AUSSI PAR LES ENTREPRISES, LES CENTRES DE RECHERCHE ET LA

POPULATION EN GÉNÉRALE. ILS SERVENT AUJOURD'HUI AUSSI BIEN À ORIENTER LES POLITIQUES DES GOUVERNEMENTS QU'À CONDITIONNER LES STRATÉGIES DE MARKETING ET À FAÇONNER L'OPINION PUBLIQUE. LORS DES ÉLECTIONS, CHAQUE GRAND RÉSEAU D'INFORMATION A SON SONDAGE ET S'EN SERT MÊME COMME OUTIL DE PROMOTION. ON PUBLIE MÊME AUJOURD'HUI LA MÉTHODOLOGIE DE SONDAGE ET LES RÉSERVES APPROPRIÉES. QUEL PROGRÈS, QUELLE RÉVOLUTION.

MAIS L'EXEMPLE DES SONDAGES D'OPINION N'EST QUE LA POINTE DE L'ICEBERG. EN EFFET, LES AUTRES CHAMPS MÉTHODOLOGIQUES ONT AUSSI ÉVOLUÉS ET LES DOMAINES STATISTIQUES SE SONT MULTIPLIÉS. L'EXPLOITATION DES FICHIERS ADMINISTRATIFS, LA SUBSTITUTION DES RECENSEMENTS PAR LES ENQUÊTES, LA PROTECTION DE LA CONFIDENTIALITÉ ET LES CONTRÔLES DE QUALITÉ EN SONT DE BONS EXEMPLES.

MAIS POURQUOI UN TEL ENGOUEMENT? LA VOLONTÉ, DEPUIS DIX ANS, POUR LES ENTREPRISES ET LES GOUVERNEMENTS DE COUPER DANS LES DÉPENSES ET/OU DE RÉDUIRE LES DÉFICITS ONT GRANDEMENT AFFECTÉ LES UNITÉS ET TRAVAUX DE RECHERCHE. IL EST BIEN CONNU QU'EN PÉRIODE DE COMPRESSIONS BUDGÉTAIRES, ON COUPE DANS LES BUDGETS DE RECHERCHE. PARALLÈLEMENT À CE PHÉNOMÈNE, IL Y A EU AUSSI CETTE VOLONTÉ DES GOUVERNEMENTS PAR EXEMPLE À RÉDUIRE LA CHARGE DE TRAVAIL POUR LES RÉPONDANTS. IRONIE DU SORT OU CONTRADICTION, LES BESOINS POUR UNE INFORMATION DE PLUS EN PLUS RAFFINÉE ET DE PLUS EN PLUS PRÉCISE SE FONT DE PLUS EN PLUS PRESSANTS.

COMMENT FAIRE PLUS ET MIEUX AVEC DE MOINS EN MOINS DE RESSOURCES TOUT EN DÉRANGEANT DE MOINS EN MOINS LES RÉPONDANTS? COMMENT AUGMENTER LA PRODUCTIVITÉ DE NOS UNITÉS DE RECHERCHE? UNE DES RÉPONSES EST L'UTILISATION DE LA MÉTHODOLOGIE STATISTIQUE DANS DE PLUS EN PLUS DE DOMAINES.

VOILÀ POURQUOI AUJOURD'HUI DANS LES AGENCES STATISTIQUES PLUTÔT QUE DE PARLER RECENSEMENTS, ON PARLE DE PLUS EN PLUS D'ENQUÊTES PAR ÉCHANTILLON OU D'EXPLOITATION DES FICHIERS ADMINISTRATIFS. VOILÀ POURQUOI DANS LES BUREAUX COMPTABLES OU LES DIRECTEURS DE CONTRÔLE DE QUALITÉ, ON NE PARLE PLUS DE VÉRIFICATION SYSTÉMATIQUE MAIS DE CONTRÔLE À PARTIR D'ÉCHANTILLONS. AVEC LES LOIS D'ACCÈS À L'INFORMATION ET LA PROTECTION DE LA DONNÉE PRIVÉE, LE RÔLE DES MÉTHODES DANS LE SECTEUR DE LA CONFIDENTIALITÉ A AUSSI PRIS BEAUCOUP D'IMPORTANCE.

IL NE FAUT PAS SE CACHER QUE LES PERFORMANCES ET LES RÉSULTATS DE MÉTHODOLOGISTES DEPUIS LES DERNIÈRES ANNÉES ONT GRANDEMENT AIDÉ LA MÉTHODOLOGIE STATISTIQUE AU POINT D'EN FAIRE UNE COMPLICE. DE PLUS EN PLUS, LES CHERCHEURS ET LES GESTIONNAIRES RECHERCHENT PLUTÔT QUE D'ÉVITER LES MÉTHODOLOGISTES PARCE QU'ILS ONT DE PLUS EN PLUS CONFIANCE EN LEURS MÉTHODES.

Participation à ce colloque

Il va sans dire que cette évolution récente a suscité des attentes vis à vis des méthodologistes et a engendré beaucoup de travaux et de développements.

L'évolution rapide du secteur d'activité force aujourd'hui les méthodologistes à se tenir à la fine pointe de cette science. En ce sens, les colloques sont des endroits privilégiés pour se mettre à jour. Les colloques sont aussi des endroits privilégiés pour présenter ses découvertes et les vérifier auprès de ses pairs. Ils permettent finalement de sortir de l'isolement des travaux quotidiens, d'échanger avec des confrères et de se créer un réseau de contacts.

Comme gestionnaire, je suis très heureux de voir que les méthodologistes plutôt que de s'asseoir dans une certaine tranquilité et de se bomber le torse suite aux succès obtenus, soient à l'affût des derniers développements et cherchent à être à la fine pointe dans leur domaine.

Le nombre de participants et de conférenciers et la nature des présentations à ce colloque sont un signe que le domaine est en pleine évolution, qu'il y a un besoin réel et qu'il y a une volonté et une soif de progrès. Tout celà est très sain, tout celà est très rafraîchissant et surtout très rassurant pour vos patrons.

Colloque de langue française

Dans un domaine où tout bouge, tous les prétextes sont bons pour multiplier les forums de discussions et d'échanges. Si la langue française en est un, pourquoi pas en profiter?

Mais il y a beaucoup plus. Un colloque comme celui ci a le net avantage de vous permettre d'échanger dans votre propre langue. La communication dans une langue autre que le français est peut être un obstacle ou un handicap pour certains d'entre vous, ce colloque règle ce genre de problème et élimine ce genre d'obstacle.

À ceux qui y verraient un problème de diversité, le fait de regrouper des méthodologistes de même langue, laisse place malgré tout à beaucoup de diversité du fait que vous vivez dans des environnements de travail très différents.

Conclusion

En conclusion, permettez-moi de souhaiter à tous les participants un colloque riche en information et des échanges des plus fructueux. J'espère que vous trouverez ici les connaissances, qui vous permettront de vous ressourcer et la volonté de tenter des expériences et de devenir des agents de changement dès votre retour dans vos milieux de travail respectifs.

PERMETTEZ-MOI FINALEMENT DE SOUHAITER AUX CONFÉRENCIERS DE TROUVER ICI UN FORUM POUR DISCUTER ET VALIDER VOS TRAVAUX. SOYEZ FIERS DU TRAVAIL QUE VOUS AVEZ ACCOMPLI ET JE SOUHAITE QUE LA VISIBILITÉ QUE VOUS FOURNIRA CE COLLOQUE VOUS MOTIVERA À CONTINUER.

BON COLLOQUE À TOUS ET MERCI POUR VOTRE PARTICIPATION.

SESSIONS SUR LA MÉTHODOLOGIE D'ENQUÊTES

UTILISATION DES TESTS ALEATOIRES DANS LE CADRE DE LA PLANIFICATION DU RECENSEMENT DE 1991

Yves Béland *

Dans le cadre de la planification du recensement de 1991, un essai a été effectué dans le but de tester les questions qui seront susceptibles d'être posées lors de ce prochain recensement. Le questionnaire de cet essai existe en deux versions: la version 'X' et la version 'Y'. La différence se situe au niveau de deux questions. Ce texte décrit une des méthodes qui a été utilisée pour comparer ces deux types de questionnaires.

MOTS CLES: Tests aléatoires, tests non-paramétriques.

1. INTRODUCTION

Dans le cadre de la planification du recensement de 1991, deux essais modulaires ont été effectués dans le but de tester les questions qui seront susceptibles d'être posées lors du prochain recensement. On veut s'assurer du libellé de chaque question pour ainsi faciliter la compréhension de celles-ci auprès des répondants.

Le premier essai qui s'est tenu au mois de novembre 1987 avait comme objectif de mettre au point de nouvelles questions portant sur la couverture, la fécondité, le travail bénévole ainsi que sur la nuptialité. Le deuxième essai, celui dont on fait mention dans ce texte et qui s'est tenu au mois de janvier 1988, visait principalement à mesurer la réaction des populations ethniques face à certaines questions portant sur l'état matrimonial, la langue, l'origine ethnique, la religion, la citoyenneté ainsi que sur la mobilité.

Le questionnaire d'essai de 1988 du deuxième essai modulaire existe en deux versions: la version 'X' et la version 'Y'. La différence se situe au niveau des questions sur l'origine et l'identité ethnique. En effet, dans la version 'X', 'CANADIEN' constitue une réponse possible (une case est réservée à cette fin) pour ces questions tandis que la version 'Y' n'offre pas ce choix de réponse pour ces mêmes questions. Pour nous permettre de tester statistiquement certaines hypothèses, on a utilisé la méthode des tests aléatoires.

* Yves Béland, Division des méthodes d'enquêtes sociales
Statistique Canada, Ottawa

2. SELECTION DE L'ECHANTILLON

Compte tenu des objectifs de l'essai modulaire, l'enquête s'est déroulée dans les six régions métropolitaines de recensement (RMR) suivantes. La région de Vancouver a été choisi à cause de son nombre de ménages formés de personnes est-asiatiques, Winnipeg pour sa population autochtone, Toronto et Montréal étant les villes les plus cosmopolites, Québec a été choisi parce qu'on voulait tester le questionnaire français et finalement Halifax pour sa population de race noire.

L'échantillon d'environ 3500 ménages du deuxième essai modulaire provient d'un échantillonnage à deux degrés. Au premier degré, un échantillon non-probabiliste de secteurs de dénombrement a été sélectionné parmi les six régions métropolitaines de recensement. Un secteur de dénombrement est une aire géographique contenant environ 250 ménages et qui constitue le plus souvent la région assignée à un énumérateur lors du recensement. Des spécialistes du sujet de la Division des statistiques sociales, du logement et des familles de Statistique Canada nous avaient fourni une liste de secteurs de dénombrement pour lesquels ils avaient la répartition des langues maternelles des ménages basée sur le recensement de 1986. Comme l'objectif principal du deuxième essai modulaire était d'observer la réaction de différentes populations ethniques face à certaines questions, on a donc sélectionné les secteurs de dénombrement où l'on croyait avoir les meilleures chances d'obtenir des ménages ayant cette caractéristique.

Au deuxième degré, un échantillon systématique de ménages a été tiré aléatoirement à partir du registre des visites du recensement de 1986 à l'intérieur de chacun des secteurs de dénombrement sélectionnés.

L'échantillon du deuxième essai modulaire n'est pas représentatif de tous les ménages du pays mais il répond bien aux objectifs de l'enquête.

3. COMPARAISON DES DEUX VERSIONS DU QUESTIONNAIRE

Comme il a été mentionné précédemment, le questionnaire du deuxième essai modulaire existe en deux versions: la version 'X' et 'Y'. Les ménages sélectionnés dans l'échantillon ont reçu aléatoirement un ou l'autre de ces questionnaires. En fait, à partir de la liste de contrôle de l'échantillon, le premier ménage se voyait assigner un type de questionnaire et le suivant recevait l'autre type et ainsi de suite. Finalement, nous avons recueilli sensiblement le même nombre de questionnaires de chaque type. Pour nous permettre de tester statistiquement certaines hypothèses, on a utilisé la méthode des tests aléatoires. La méthode des tests aléatoires peut être utilisée pour comparer deux traitements appliqués aux unités d'échantillons qui ne sont pas nécessairement probabilistes.

On désirait savoir si les questions du questionnaire d'essai, portant sur l'identité et l'origine ethnique du questionnaire d'essai, amènent

plus ou moins de réponses multiples dans la version 'X' que dans la version 'Y'. On dit qu'il y a une réponse multiple lorsque deux cases ou plus sont cochées par un répondant. On voulait aussi connaitre l'influence du type de questionnaire sur la multiplicité (nombre de cases cochées) et sur certains choix de réponse (comme FRANÇAIS) pour ces questions. Etant donné les différentes populations cibles associées aux régions métropolitaines de recensement lors de la sélection de l'échantillon, ces tests aléatoires ont été faits séparément pour chacune de ces régions.

Tout d'abord, pour une question donnée, on calcule la proportion de questionnaires avec une réponse multiple pour chacune des versions 'X' et 'Y' qu'on dénote P_x et P_y. Par la suite, pour simuler ce qui se passerait si le type de questionnaire n'avait aucun effet, on permute, parmi les ménages, les types de questionnaire et on calcule les nouvelles valeurs de P_x et P_y. Idéalement, si le temps et l'argent nous le permettait, on calculerait les valeurs de P_x et P_y pour chacune des permutations possibles, mais ces dernières sont beaucoup trop nombreuses. C'est la raison pour laquelle on sélectionne un échantillon de 999 permutations. Un fait à noter, on permute les types de questionnaire de façon à ce que les personnes d'un même ménage se voient assigner le même type de questionnaire.

Ensuite, pour chacune des 999 permutations de types de questionnaire ainsi que pour l'échantillon initial observé, on calcule la statistique $s=P_x-P_y$. On se retrouve donc avec 1000 valeurs de s que l'on classe en ordre croissant. Si, lors de cette classification, deux ou plusieurs statistiques sont égales, un nombre aléatoire entre 0 et 1 est généré et c'est ce nombre qui déterminera l'ordre parmi les statistiques égales. La variable $RANGP_{x-y}$ représente le rang de la statistique s observée.

Statistiquement parlant, pour toutes les RMR, excluant celle de Halifax, on teste l'hypothèse suivante:

> H_0: L'espérance de la proportion de questionnaires avec une réponse multiple pour la version 'X' est la même que pour la version 'Y'.

versus

> H_1: L'espérance de la proportion de questionnaires avec une réponse multiple est plus 'grande' pour la version 'X' que pour la version 'Y'.

et on rejette H_0 si le niveau critique est inférieur à 0.05. Pour Halifax, on teste l'hypothèse:

> H_0: L'espérance de la proportion de questionnaires avec une réponse multiple pour la version 'X' est la même que pour la version 'Y'.

versus

H_1: L'espérance de la proportion de questionnaires avec une réponse multiple est plus 'petite' pour la version 'X' que pour la version 'Y'.

et on rejette H_0 si le niveau critique est inférieur à 0.05.

La raison pour laquelle on utilise une contre-hypothèse (H_1) différente pour Halifax comparativement aux autres RMR est que la majorité des ménages qui ont été sélectionnés dans cette RMR sont composés de personnes de souche britannique. Dans le questionnaire de type 'Y', 'CANADIEN' n'est pas un choix de réponse pour les questions sur l'identité et l'origine ethnique, et comme la plupart des personnes de cette RMR sont d'origine anglaise, écossaise ou irlandaise, ces gens qui ont reçu un questionnaire 'Y' ont coché une ou plusieurs de ces cases. Pour ce qui est des ménages qui ont reçu un questionnaire 'X', les personnes de ces ménages peuvent tout simplement coché la case 'CANADIEN'. Comme on s'attend à avoir plus de réponses multiples pour le type 'Y', cette contre-hypothèse s'impose.

Le niveau critique est calculé de la manière suivante: pour la RMR de Halifax, étant donné qu'on rejette H_0 si la proportion chez les 'X' est significativement inférieure à celle observée chez les 'Y', le niveau critique se défini comme $RANGP_{x-y}/1000$. Par contre, pour toutes les autres RMR, étant donné qu'on rejette H_0 si la proportion chez les 'X' est significativement supérieure à celle observée chez les 'Y', le niveau critique se défini comme $(1001-RANGP_{x-y})/1000$.

Prenons un exemple concret. A Winnipeg, on a 373 ménages dont 187 ont reçu un questionnaire de type 'X'. On a évalué, pour la question sur l'origine ethnique, la proportion de questionnaires avec une réponse multiple: P_x=0.293 et P_y=0.307. Un échantillon de 999 permutations de type de questionnaire a été sélectionné et les 999 nouvelles valeurs de P_x et P_y, associées aux permutations, ont été calculées. Par la suite, la statistique $s=P_x-P_y$ a été calculée pour les 1000 ensembles de données (999 associés à l'échantillon de permutations + celui de la valeur observée). Ces 1000 statistiques s ont été classées en ordre croissant. Le rang de la statistique observée, $RANGP_{x-y}$, est 379 ce qui procure un niveau critique de 0.622. On conclut donc qu'il n'y a pas de différence significative entre les deux types de questionnaire quant au nombre de réponses multiples car 0.622>0.050.

Regardons ce qui se passe du côté de Québec pour la question sur l'origine ethnique. Comme $RANGP_{x-y}$=1000, on obtient un niveau critique de 0.001 et donc il y a effectivement une différence significative. En effet, il y a beaucoup plus de réponses multiples dans le questionnaire 'X' que dans le questionnaire 'Y' car plusieurs gens ont répondu 'CANADIEN' et 'FRANCAIS' plutôt que 'FRANCAIS'. Cependant, pour la question sur l'identité ethnique à Halifax, on observe la situation inverse. En effet, la proportion de questionnaires avec une réponse multiple est significativement inférieure pour la version 'X' que pour la version 'Y'. Enfin, les valeurs de $RANGP_{x-y}$ qui sont significatives pour les différentes RMR sont marquées d'un astérisque dans les tableaux fournis en annexe à la fin de ce texte.

La méthode des tests aléatoires a également été utilisée pour tester la multiplicité (nombre de cases cochées) des questions sur l'origine et l'identité ethnique pour chacune des RMR. Cette fois-ci, au lieu de parler de proportions (P_x, P_y), on parle de ratios (R_x, R_y). Le ratio R_x (R_y) est le nombre moyen de cases cochées pour une question des questionnaires 'X' ('Y'). Par la suite, la méthode demeure la même et au lieu de parler de $RANGP_{x-y}$, on parle de $RANGR_{x-y}$ et la statistique s se définit maintenant comme étant R_x-R_y. Cependant, étant donné la plus grande variabilité des statistiques s pour les tests quant à la multiplicité, on a pris un échantillon de 1999 permutations au lieu de 999. Les 2000 statistiques ont été classées de la même manière que précédemment mais pour demeurer homogène dans les tableaux, les $RANGR_{x-y}$ ont été divisés par deux. Comme pour les tests pour les réponses multiples, on considère qu'il y a une différence significative si le niveau critique ($RANGR_{x-y}/1000$ ou $(1001-RANGR_{x-y})/1000$) est inférieure à 0.05.

Pour toutes les RMR, excluant celle de Halifax, on teste l'hypothèse suivante:

H_0': L'espérance du nombre moyen de cases cochées pour la version 'X' est la même que pour la version 'Y'.

versus

H_1': L'espérance du nombre moyen de cases cochées est plus élevé pour la version 'X' que pour la version 'Y'.

et on rejette H_0' si le niveau critique est inférieur à 0.05. Pour Halifax, on teste l'hypothèse:

H_0': L'espérance du nombre moyen de cases cochées pour la version 'X' est la même que pour la version 'Y'.

versus

H_1': L'espérance du nombre moyen de cases cochées est moins élevé pour la version 'X' que pour la version 'Y'.

et on rejette H_0' si le niveau critique est inférieur à 0.05. Les valeurs de $RANGR_{x-y}$ qui sont significatives sont marquées d'un astérisque dans les tableaux.

Pour terminer, on a comparé les versions 'X' et 'Y' du deuxième essai modulaire quant au nombre de Français, Britanniques et Sud-Asiatiques dans certaines RMR. On a considéré comme Britannique quiconque a coché une des cases 'IRLANDAIS', 'ECOSSAIS' ou 'ANGLAIS'. Comme pour les tests sur le nombre de réponses multiples, P_x (P_y) est la proportion de questionnaires 'X' ('Y') qui ont répondu Français (Britannique ou Sud-Asiatique selon le cas) dans une RMR. On voulait savoir si la proportion de Français, Britanniques et Sud-Asiatiques chez les

questionnaires 'X' est significativement inférieure à celle observée chez les questionnaires 'Y'. Pour ce faire, on a utilisé la méthode des tests aléatoires avec un échantillon de 999 permutations.

Pour toutes les RMR, on teste l'hypothèse suivante:

H_0'': L'espérance de la proportion de gens qui ont répondu Français (Britannique ou Sud-Asiatique) est la même pour la version 'X' que pour la version 'Y'.

versus

H_1'': L'espérance de la proportion de gens qui ont répondu Français (Britannique ou Sud-Asiatique) est plus 'petite' pour la version 'X' que pour la version 'Y'.

et on rejette H_0'' si le niveau critique est inférieur à 0.05. Etant donné qu'on rejette H_0'' si la proportion chez les 'X' est significativement inférieure à celle observée chez les 'Y', le niveau critique se défini comme $RANGP_{x-y}/1000$. Les valeurs de $RANGP_{x-y}$ qui sont significatives sont marquées d'un astéristique dans les tableaux.

4. ANNEXE

QUESTION SUR L'ORIGINE ETHNIQUE						
RMR	MULTIPLE			MULTIPLICITE		
	Px	Py	RANGPx-y	Rx	Ry	RANGRx-y
HALIFAX	0.435	0.536	87	1.617	1.914	62
QUEBEC	0.154	0.043	1000*	1.143	0.986	1000*
MONTREAL	0.185	0.194	389	1.141	1.152	416
TORONTO	0.127	0.122	608	1.124	1.125	506
WINNIPEG	0.293	0.307	379	1.439	1.398	656
VANCOUVER	0.285	0.296	380	1.440	1.392	721

QUESTION SUR L'IDENTITE ETHNIQUE						
RMR	MULTIPLE			MULTIPLICITE		
	Px	Py	RANGPx-y	Rx	Ry	RANGRx-y
HALIFAX	0.220	0.335	35*	1.244	1.502	29*
QUEBEC	0.140	0.016	1000*	1.131	0.959	1000*
MONTREAL	0.159	0.125	938	1.075	1.044	815
TORONTO	0.186	0.120	1000*	1.154	1.075	996*
WINNIPEG	0.224	0.195	753	1.253	1.208	703
VANCOUVER	0.186	0.183	544	1.182	1.137	799

QUESTION SUR L'ORIGINE ETHNIQUE				
VARIABLE	RMR	Px	Py	RANGPx-y
FRANCAIS	QUEBEC	0.127	0.897	1*
FRANCAIS	MONTREAL	0.038	0.210	1*
BRITANNIQUE	HALIFAX	0.321	0.837	1*
BRITANNIQUE	MONTREAL	0.034	0.092	2*
BRITANNIQUE	TORONTO	0.085	0.135	3*
BRITANNIQUE	WINNIPEG	0.167	0.234	54
BRITANNIQUE	VANCOUVER	0.267	0.325	65

QUESTION SUR L'IDENTITE ETHNIQUE				
VARIABLE	RMR	Px	Py	RANGPx-y
FRANCAIS	QUEBEC	0.138	0.899	1*
FRANCAIS	MONTREAL	0.034	0.217	1*
BRITANNIQUE	HALIFAX	0.153	0.828	1*
BRITANNIQUE	MONTREAL	0.022	0.117	1*
BRITANNIQUE	TORONTO	0.050	0.215	1*
BRITANNIQUE	WINNIPEG	0.074	0.276	1*
BRITANNIQUE	VANCOUVER	0.104	0.325	1*
ITALIEN	TORONTO	0.412	0.463	60
SUD-ASIATIQUE	TORONTO	0.087	0.055	962
SUD-ASIATIQUE	VANCOUVER	0.259	0.238	638

La pondération lors du recensement de la population de 1986

par Pierre Daoust*

Résumé

Lors du recensement de la population du Canada de 1986, de l'information supplémentaire a été receuillie pour un échantillon représentant un cinquième des ménages privés. La stratégie de pondération utilisée pour produire les estimations des caractéristiques échantillonnées était presque la même que celle utilisée lors du recensement de 1981. Une évaluation plus poussée a permis d'obtenir une banque d'information de meilleur qualité sur le processus de pondération. L'information plus détaillée obtenue pour le recensement de 1986 en rapport aux problèmes rencontrés, permettra de mieux nous orienter lors de la ré-évaluation de la stratégie de pondération prévue pour le recensement de 1991.

Mot Clef: Méthode itérative d'estimation par le quotient

INTRODUCTION

Le jour du recensement en 1986, une profusion de renseignements ont été recueillis auprès de tous les habitants du Canada. Neuf questions portant sur l'âge, le sexe, l'état matrimonial, la langue maternelle, le lien avec la personne repère du ménage (Personne 1), le type de logement et le mode d'occupation ont été posées à tous les canadiens. Cette information est appelée l'information 2A. Des renseignements additionnels ont été obtenus pour un échantillon formé de 20% des logements privés occupés et la majorité des logements collectifs. Les canadiens dans l'échantillon ont répondus aux neuf questions mentionnées ci-dessus, à quatre questions relatives au logement et 19 autres questions portant sur des aspects socio-économiques. Ces questions ont été posées à un échantillon de répondants afin de diminuer les coûts et d'alléger le fardeau de déclaration. Cette information échantillonnée est connue comme l'information 2B.

* Pierre Daoust est méthodologiste à la division des méthodes d'enquêtes sociales à Statistique Canada, Ottawa(Ontario), K1A 0T6

Pour obtenir les estimations basés sur l'échantillon qui apparaissent dans les publications du recensement, on utilise une quantité appelée poids (aussi appelée coefficients de pondération). Des poids distincts sont déterminés pour chaque ménage et pour chaque personne dans l'échantillon. Les poids des personnes servent également à déterminer les poids attribués aux familles. Des données dites pondérées sont par la suite obtenues en multipliant ces poids aux caractéristiques de l'échantillon. Les estimations sont finalement obtenues en additionnant les données pondérées des unités correspondentes (ménages, personnes, familles).

La valeur des poids est d'abord fixée à 5 dans les régions où on a choisi un échantillon au taux d'un logement privé occupé sur 5. On procède de la sorte parce que tout ménage échantillonné doit non seulement se représenter lui-même, mais doit aussi représenter quatre ménages qui ne font pas partie de l'échantillon. Le poids de 5 est de plus ajusté afin que les estimations finales (basées sur l'échantillon) concordent avec certains chiffres basés sur une énumération complète de la population (soit les caractéristiques 2A).

Cette technique qui consiste à utiliser des poids ajustés est appelée méthode itérative d'estimation par le quotient. Elle permet d'assurer que les chiffres du recensement selon certaines caractéristiques 2A (par exemple l'âge, le sexe et l'état matrimonial) dans une région géographique sont identiques, peut importe si l'on utilise des données recueillies auprès de la population totale ou des données pondérées provenant de l'échantillon. Les estimations basées sur ces poids ajustés sont aussi plus précis que celles basées sur les poids simples de valeur 5.

Suite au recensement de 1986, une évaluation du processus de pondération a été produite. Cette évaluation visait principalement l'analyse des concordances, au niveau des provinces et du canada, entre certains chiffres produits à l'aide de la population totale et ceux correspondants produits à l'aide des données pondérées. Les caractéristiques étudiées étaient celles utilisées par la stratégie de pondération en vue de l'ajustement des poids. A cette fin plusieurs propriétés de ces caractéristiques ainsi que les éléments influencant la performance de la méthode itérative par le quotient furent étudiées. L'analyse graphique effectuée lors de cette évaluation a permis de bien identifier les avantages et inconvénients de la stratégie de pondération actuelle. L'information receuillie s'est avérée d'un précieux apport pour la ré-évaluation de la stratégie de pondération prévue pour le recensement de 1991.

La première section de ce rapport présente un sommaire de la stratégie de pondération telle qu'utilisée en 1986. Les fait saillants de l'évaluation effectuée suite au recensement sont mentionnés dans la deuxième section. Finalement les recommendations pour le recensement de 1991 sont incluses dans la troisième section.

1. STRATÉGIE DE PONDÉRATION POUR 1986

Cette section présente un sommaire de la stratégie de pondération utilisée lors du recensement de 1986. Ce sommaire, ainsi qu'une partie de l'introduction, est dérivé du bulletin d'information à l'intention des utilisateurs 'Méthodes utilisées pour établir des estimations à partir de l'échantillon du recensement' écrit par Michael Bankier, méthodologiste principal de la Division des Méthodes d'Enquêtes Sociales. Une description détaillée de la stratégie de pondération peut être obtenue dans Rakhra (1986).

1.1 CHOIX DE L'ÉCHANTILLON DU RECENSEMENT

Voici comment l'échantillon utilisé pour le recensement a été choisi:

a) Dans les secteurs de dénombrement (SD) où le recensement s'est fait par autodénombrement chaque cinquième logement privé occupé a été choisi pour faire partie de l'échantillon.

b) Dans les secteurs de recensement par interview (généralement les régions éloignées ou les réserves indiennes) tous les logements privés occupés ont été sélectionnés dans l'échantillon.

c) On a posé à toutes les personnes qui habitent des logements collectifs* ainsi qu'aux canadiens en poste à l'étranger (généralement le personnel des ambassades ou les membres des Forces armées) les mêmes questions qu'aux personnes qui font partie de l'échantillon à l'exclusion des questions sur le logement. Cependant, les questions relatives au lieu habituel de résidence de la personne au Canada ont été posées.

* Les résidents de logements collectifs suivants n'ont pas répondus aux questions supplémentaires:

(i) détenus dans les établissements de correction et les établissements pénitentiaires ou les prisons;

(ii) patients dans les hôpitaux généraux, dans les centres de soins spéciaux et dans les établissements pour personnes âgées et pour les malades chroniques ou dans les hôpitaux psychiatriques;

(iii) enfants dans les orphelinats et les foyers pour enfants ou les foyers pour jeunes contrevenants.

1.2 DÉTERMINATION DES POIDS UTILISÉS POUR LE RECENSEMENT

Le calcul des poids utilisés pour le recensement se fait en quatre étapes. Tout d'abord, des secteurs de dénombrement (SD) sont combinés pour former des régions de pondération (RP). Une RP correspond au plus petit niveau géographique pour lequel on peut garantir qu'il y a concordance entre certaines valeurs estimées et des chiffres connus relatifs à la population. Les renseignements portant sur la population et sur l'échantillon, pour la RP, font alors l'objet d'un classement recoupé dans un tableau dont les lignes et les colonnes sont définies à l'aide de caractéristiques connues pour l'ensemble de la population (par exemple, l'âge selon le sexe). Ces lignes et colonnes sont ensuites combinées, au besoin, afin que l'échantillon et la population soient assez considérables pour que l'algorithme de pondération puisse bien fonctionner. Finalement, les poids sont déterminés de façon à ce que le nombre estimé de personnes ou de ménages qui se trouve dans chaque ligne ou dans chaque colonne corresponde aux chiffres connus pour toute la population. Ces étapes sont décrites plus en détail ci-dessous.

1.2.1 Formation des régions de pondération

Des SD contigus sont groupés pour former des régions de pondération(RP). Une RP contient généralement entre 2,000 et 7,000 personnes dans des ménages privés qui font partie de SD pour lesquels un échantillon d'un ménage privé sur 5 a été choisi.

Il ne faut pas qu'une RP chevauche les limites des divisions de recensement et, dans la mesure du possible, elle doit respecter celles des circonscriptions électorales fédérales, des subdivisions de recensement et des secteurs de recensement. Tous les calculs qui restent à faire en rapport avec les poids sont effectués indépendamment pour chacune des 5,347 RP au Canada.

1.2.2 Classement recoupé des ménages et des personnes

On définit un tableau de classement recoupé ("c r") pour les ménages et un tableau "c r" pour les personnes au niveau de la RP d'après les caractéristiques connues pour l'ensemble de la population. Les lignes et les colonnes de ces tableaux sont décrites aux figures 1 à 4 de l'annexe I. On détermine le nombre de ménages dans la population ainsi que dans l'échantillon qui se trouve dans les diverses cellules du tableau "c r" pour les ménages. De même, on détermine le nombre de personnes dans la population et dans l'échantillon qui se trouve dans les diverses cellules du tableau "c r" pour les personnes. Seuls les ménages privés des SD où on a choisi un échantillon de 20% des ménages ainsi que les personnes qui habitent ces ménages sont compris dans les chiffres relatifs à la population et à l'échantillon.

1.2.3 Combinaison des données contenues dans les tableaux de classement recoupé

Les lignes des tableaux "c r" sont combinées avec d'autres lignes si certains critères relatifs aux totaux des lignes pour la population et pour l'échantillon ne sont pas respectés. Plus précisement, une ligne est combinée à une autre si:

a) le total relatif à la population pour la ligne est inférieur à 35;

b) le total relatif à l'échantillon pour la ligne est 0;

c) le rapport entre le total relatif à la population pour la ligne et le total relatif à l'échantillon pour cette même ligne n'est pas compris entre 3.0 et 19.9.

Les stratégies utilisées pour combiner les lignes des tableaux "c r" sont également données aux figures 1 à 4 de l'annexe I. Quelques points utiles pour interpréter ces figures doivent être mentionnés. La figure 1, représentant les lignes du tableau "c r" des ménages, est utilisée à titre d'exemple.

a) Les lignes doubles délimitent les groupes de lignes pouvant être combinées entre elles. Par exemple, les lignes 1 à 6 peuvent être combinées, tout comme les lignes 7 à 14. Toutefois, les lignes 6 et 7, parce qu'elles appartiennent à deux groupes différents, ne peuvent être combinées à moins que les stratégies de combinaison ne soient épuisées.

b) Les lignes d'un même groupe sont numérotées (voir la colonne "étiquette") consécutivement en débutant avec le chiffre un. Par exemple, les lignes 7 à 14 sont numérotées de 1 à 8. Ceci a été fait afin d'éviter, par simplicité, les nombres de deux chiffres lors de la définition des stratégies de regroupement.

c) Parmi les colonnes identifiées "Stratégie de combinaison" le nombre "4365", par exemple, indique que les lignes numérotées 4, 3, 6 et 5 de ce groupe devraient être combinées. La ligne combinée est alors nommée 4365. Pour le groupe représenté par les lignes de 7 à 14, 4365 indique que les lignes 10, 9, 12 et 11 devraient être combinées, parce que la ligne 10 a initialement échouée aux critères mentionnés et a dû être combinée.

Lorsque les stratégies de combinaison sont épuisées et qu'il existe une ligne pour laquelle le total relatif à l'échantillon est nulle mais dont la total relatif à la population est différent de 0, des combinaisons supplémentaires sont nécessaires. Cette ligne est alors combinée avec la ligne ayant le plus petit rapport entre le total relatif à la population et le total relatif à l'échantillon.

Les critères de combinaison des colonnes des tableaux "c r" sont identiques à ceux mentionnés pour les lignes. Les stratégies de combinaison sont données aux figures 2 et 4 de l'annexe I.

1.2.4 Calcul des poids

Des poids de 1 sont attribués aux ménages collectifs ainsi qu'aux ménages privés dans les SD éloignés, sur les réserves indiennes et à l'étranger puisque, dans ces cas, on n'a pas eu recours à l'échantillonnage. Les personnes qui habitent ces ménages se voient aussi attribuer un poids de 1. On attribue un poids initial de 5 à toutes les personnes et à tous les ménages dans l'échantillon. Pour chaque colonne, le total estimé pour la colonne obtenu à l'aide de ce poids initial est comparé au total connu de la population pour cette colonne. On élimine toute divergence au niveau de la colonne en multipliant le poids initial par le rapport entre la population totale de la colonne et le total estimé pour la colonne.

Ce poids révisé est appelé le poids de première itération. Puis, pour chaque ligne, le total estimé pour la ligne obtenu à l'aide de ce poids de première itération est comparé au total connu de la population pour la ligne. On élimine toute divergence au niveau des lignes en multipliant le poids de première itération par le rapport entre la population totale pour la ligne et le total estimé pour la ligne. Ce poids révisé est appelé le poids de deuxième itération. Ce processus d'élimination des divergences dans les estimations pour les colonnes et pour les lignes se poursuit jusqu'à ce que toutes les divergences qui restent soient négligeables. Pour être spécifique, le processus est arrêté si après une itération sur les lignes la "mesure de convergence" pour les colonnes donnée par

$$\frac{\sqrt{\sum_j \left(\left(\begin{array}{c}\text{total estimé de}\\ \text{la colonne } j\end{array}\right) - \left(\begin{array}{c}\text{total de population}\\ \text{de la colonne } j\end{array}\right)\right)^2}}{\sqrt{\sum_j \left(\begin{array}{c}\text{total de population}\\ \text{de la colonne } j\end{array}\right)^2}}$$

n'a pas diminué de plus de 0.01 depuis la dernière mesure calculée. Le processus ne peut être terminé avant qu'un minimum de quatre itérations n'aient été complétées et est automatiquement terminé après 80 itérations. Ces poids finaux pour les ménages et pour les personnes sont alors utilisés pour produire les estimations officielles du recensement.

C'est le poids attribué au mari ou au parent seul qui est utilisé comme poids pour la famille de recensement. Similairement, c'est le poids attribué au mari de la personne repère de la famille économique ou, si le mari n'est pas présent, celui attribué à la personne repère de la famille économique qui est utilisé comme poids pour la famille économique. Les définitions de familles de recensement et de familles économiques sont données à l'annexe II.

Pour de plus amples renseignements sur la méthode itérative d'estimation par le quotient, voir la bibliographie à la section 4

2. SOMMAIRE DE L'ÉVALUATION SUR LA PONDÉRATION

(a) Un des objectifs de la pondération est de minimiser la différence entre les chiffres basés sur une énumération complète définis par les lignes et les colonnes des tableaux "c r" et leurs estimations respectives basées sur un échantillon de la population. Les figures 5a et 5b de l'annexe I présentent les pourcentages de différences population/échantillon au niveau régional (l'Est, le Québec, l'Ontario et l'Ouest, cette dernière région incluant le Yukon et les Territoires du Nord Ouest) et au niveau du Canada pour les lignes et les colonnes du tableau "c r" des ménages. La différence population/échantillon est définie par l'équation [(estimation -chiffre de population)/ chiffre de population]. Les figures 6a et 6b présentent des résultats similaires pour les lignes et les colonnes du tableau "c r" pour les personnes. Nous voyons que, sauf quelques exceptions près, les différences sont relativement faibles. Les lignes et les colonnes dont les différences dépassent 1% sont habituellement petites (c.à.d. leurs totaux relatifs à la population étaient inférieurs à 35 au niveau des RP) et de ce fait ont été combinées fréquemment avec des lignes et des colonnes plus considérables. Les différences sont souvent reliées à des variations importantes entre les pourcentages échantillonnés des populations pour ces petites lignes ou petites colonnes en comparaison aux pourcentages des lignes ou des colonnes plus considérables. Les causes de ces variations des fractions d'échantillonnage ne sont pas complètement identifiées. Elles peuvent être dues à la variation introduite par l'échantillonnage ou encore par un échantillon biaisé. L'échantillon peut être biaisé par les actions des énumérateurs, des répondants ou encore lors du traitement des données à Statistique Canada. Une étude sera effectuée afin de déterminer la portée de ce biais. Il a déjà été démontré, voir Daoust (1988), qu'une sous-estimation de la catégorie "autre" de la langue maternelle est due à une plus grande proportion de répondants 2A ayant rapportés une réponse multiple à la question sur la langue maternelle comparativement à la proportion observée chez les répondants 2B.

(b) Une comparaison avec les résultats obtenus en 1981 nous permet d'observer une diminution des fractions d'échantillonnage pour plusieurs lignes et colonnes des tableaux "c r" ainsi qu'une plus grande variabilité des fractions d'échantillonnage. Ceci peut être causé par une augmentation des conversions de questionnaires 2B en questionnaires 2A en 1986 comparativement à 1981(voir Carter (1987)). Ces conversions ont pour but de corriger des cas de non-réponse totale des questions échantillonnées. Elles sont effectuées afin de corriger par la pondération et non par imputation le problème de non-réponse totale. La plus grande variation parmi les fractions d'échantillonage peut causer, tel que décrit en (a), des différences population/échantillon plus importantes que celles obtenus en 1981.

(c) Certaines des différences population/échantillon pour les colonnes du tableau "c r" des personnes ont été causées par un problème de nonconvergence de la méthode itérative d'estimation par le quotient. Ceci peut s'être produit parce qu'il était impossible d'obtenir la cohérence population/échantillon pour toutes les lignes et toutes les colonnes de la matrice "c r" (celles-çi sont alors connues commes des contraintes incohérentes). L'utilisation de l'âge pour définir à la fois les lignes et les colonnes de la matrice "c r" des personnes a produit des contraintes incohérentes pour certaines RP. Ceci est la cause probable d'une surestimation inattendue des enfants âgés de 0 à 14 ans. Dans certains autres cas, la convergence peut ne pas être obtenue parce que le processus itératif a été interrompu trop rapidement. Dans un cas particulier, il a été interrompu après quatre itérations parce que la mesure de convergence avait augmentée au lieu de diminuer comme prévu.

(d) Le total estimé du nombre de familles de recensement en 1986 était de 1133 inférieur au nombre actuel connu au niveau du Canada. En comparaison, une surestimation de 337 familles avait été reportée en 1981. L'incohérence, bien que plus importante qu'en 1981, ne représente que 0.0169% du nombre total de familles de recensement comparativement à 0.0053% en 1981 . Puisque l'estimation du nombre de familles de recensement est contrôlée par les colonnes 1 à 9 de la matrice "c r" des personnes (et qu'aucune combinaison n'est survenue avec d'autres colonnes) cette incohérence est le résultat de la non-convergence du processus de pondération. D'autre part, le total estimé de familles économiques en 1986 était de 1577 plus élevé que le nombre connu au niveau du Canada. Ceci représente une amélioration considérable comparativement à une surestimation de 3738 en 1981. Cette différence ne représente toutefois que 0.0232% du nombre total de familles économiques comparativement à 0.0586% en 1981. Les familles économiques ne sont pas contrôlées de façon explicite par la stratégie de pondération.

(e) La valeur absolue des différences population/échantillon pour les subdivisions du recensement(SDR) ayant une population supérieur à 3400 n'a pas excédé 3%. Pour obtenir de tel résultats, la création de certaines RP ne contenant qu'un seul SD a été nécessaire. Ceci a été fait pour 2 SD ayant des fractions d'échantillonnage près de 1 et pour 3 SD ayant des fractions d'échantillonnage près de 0. Il s'avèra nécessaire de procéder ainsi puisque un poids initial constant est appliqué à tous les SD d'une RP. En conséquence, les SD ayant des fractions d'échantillonnage beaucoup plus petites que la moyenne des SD d'une RP ont des estimations biaisées négativement alors que les SD ayant des fractions d'échantillonnage beaucoup plus hautes que la moyenne ont des estimations biaisées positivement.

(f) Approximativement 95% des poids des ménages et des personnes sont dans l'intervalle 3-7, ce qui semble raisonnable. Étrangement, toutefois, les modes des distributions des poids des ménages et des personnes sont dans l'intervalle 4.00-4.99 bien que l'inverse de la fraction d'échantillonage globale(et celles des lignes et colonnes des matrices "c r") soit de plus de 5 (5.14 pour les ménages et 5.12 pour les personnes). Ceci peut être causé par la forme de l'équation minimisée par la méthode itérative d'estimation par le quotient.

(g) Les SD des réserves indiennes ayant refusées de répondre au recensement ont été exclus du calcul des poids. Lors du recensement de 1986 seul des estimations provinciales ont été produites pour cette population. Ces ménages et ces personnes ne sont pas considérés pour les tableaux détaillés produits par le recensement.

3. RECOMMENDATIONS POUR 1991

(a) Une étude sera prochainement entreprise afin de déterminer les causes de la variabilité des fractions d'échantillonnage pour diverses sous-populations du recensement. En particulier, l'impact des conversions 2B/2A pour le traitement de la non-réponse totale sera étudié très sérieusement. Bankier (1988a) présente une stratégie possible pour la conversion 2B/2A. De plus, toute évidence d'un biais dans l'échantillon dû à l'énumérateur ou aux répondants devra être établie. Des recommendations seront produites pour s'assurer que les procédures sur le terrain, au traitement et à la pondération restreignent adéquatement la variabilité des fractions d'échantillonage.

(b) La combinaison des lignes ou des colonnes ayant des fractions d'échantillonnage semblables devrait être considérée, lorsque possible, afin de réduire les différences population/échantillon. D'autres modifications possibles aux stratégies de combinaison sont présentées dans Bankier (1988b).

(c) La variable langue maternelle ne devrait pas être utilisée comme variable de classement recoupé. Cette variable est souvent la cause de combinaison et a donc un usage limité. L'utilisation de techniques d'estimation plus générales que la méthode itérative d'estimation par le quotient pourrait possiblement permettre de garantir la cohérence population/échantillon pour la langue maternelle sans que celle-çi ne soit utilisée comme variable de classement recoupé.

(d) La variable âge ne devrait pas être utilisée pour définir les colonnes de la matrice "c r" des personnes parce qu'elle crée des contraintes incohérentes pour le processus itératif. Éliminer cette variable peut permettre également de réduire les différences population/échantillon pour les familles de recensement.

(e) Le critère d'arrêt pour le processus de pondération itératif devrait être ré-évalué. Dans certains cas, le processus itératif s'est terminé trop tôt. En conséquence, un minimum de dix itérations devraient être effectuées avant de pouvoir terminer le processus. Dans d'autres cas (avec la matrice "c r" des personnes) jusqu'à 80 itérations on été effectuées. Le gain obtenu après autant d'itérations devrait être étudié.

(f) Des changements significatifs aux définitions des lignes et des colonnes de la matrice "c r" des ménages pourraient être nécessaires si la proposition de permettre d'identifier plus d'une personne soutient par ménage est implantée en 1991.

(g) Les poids initiaux devraient correspondre à l'inverse des fractions d'échantillonnage réalisées des SD. En 1986, un poids initial constant était appliqué aux ménages et personnes de tous les SD d'une RP. Utiliser les fractions d'échantillonnage pour déterminer les poids initiaux rendrait probablement inutile la création de RP contenant un seul SD.

(h) Les avantages et désavantages de la méthode itérative d'estimation par le quotient devraient être comparés à ceux d'estimateurs créés depuis que la méthode a été implantée lors du recensement de 1971 qui pourraient satisfaire aux objectifs visés par la pondération du recensement.

4. RÉFÉRENCES

Bankier, M.D. 1986. Estimators Based on Several Stratified Samples With Applications to Multiple Frame Survey. Journal of the American Statistical Association. Vol. 81, no. 396, 1074-1079.

Bankier, M.D. 1988a. Strategy For 1991 Census 2A/2B Conversions. Note de service de la Division des méthodes d'enquêtes sociales, 15 novembre.

Bankier, M.D. 1988b. Post-Stratification of the Census Sample. Rapport interne de la Division des méthodes d'enquêtes sociales.

Brackstone, G.J. et Rao, J.N.K. 1979, An Investigation of Raking Ratio Estimators. Sankhya. Série C, partie 2, 41, 97-114.

Daoust, P. 1988, Analysis of Large 2A/2B Inconsistency for Multiple Responses to Mother Tongue Variable, Note de service de la Division des méthodes d'enquêtes sociales, 18 mars.

Rakhra, M. 1986. 1986 Weighting Characteristics Requirements. Rapport interne de la Division des méthodes d'enquêtes sociales.

ANNEXE I

Figure 1: Lignes du tableau de classement recoupé pour les ménages

Ménage	Soutien du ménage		Nombre de personnes dans le ménage	No de ligne	étiquette	Stratégie de combinaison						
	Sexe	Âge				étape 1	étape 2	étape 3	étape 4	étape 5	étape 6	étape 7
Ménages composés d'une famille ou plus	Masculin	15-24	=2	1	1	12	123	1234	12345	123456		
			>2	2	2	21	214	2143	21436	214365		
		25-34	=2	3	3	34	341	3412	34125	341256		
			>2	4	4	43	432	4321	43216	432165		
	Féminin	15-34	=2	5	5	56	561	5612	56123	561234		
			>2	6	6	65	652	6521	65214	652143		
	Masculin	35-44	=2	7	1	12	123	1234	12345	123456	1234567	12345678
			>2	8	2	21	214	2143	21436	214365	2143658	21436587
		45-54	=2	9	3	34	345	3456	34561	345612	3456127	34561278
			>2	10	4	43	436	4365	43652	436521	4365218	43652187
		55-64	=2	11	5	56	563	5634	56341	563412	5634127	56341278
			>2	12	6	65	654	6543	65432	654321	6543218	65432187
	Féminin	35-64	=2	13	7	78	785	7856	78563	785634	7856341	78563412
			>2	14	8	87	876	8765	87654	876543	8765432	87654321
	Masculin	≥65	=2	15	1	12	123					
			>2	16	2	21	213					
	Féminin	≥65		17	3	31	312					
Ménages non familiaux composés d'une personne	Masculin	15-34		18	1	12	123	1234	12345	123456		
	Féminin	15-34		19	2	21	214	2143	21436	214365		
	Masculin	35-64		20	3	34	341	3412	34125	341256		
	Féminin	35-64		21	4	43	432	4321	43216	432165		
	Masculin	≥ 65		22	5	56	563	5634	56341	563412		
	Féminin	≥ 65		23	6	65	654	6543	65432	654321		
Ménages non familiaux composés de deux personnes ou plus	Masculin			24	1	12						
	Féminin			25	2	21						

Figure 2: Colonnes du tableau de classement recoupé pour les ménages

Mode d'occupation du logement	Type de logement	Nº de colonne	étiquette	Stratégie de combinaison
				étape 1
Logement possédé	Logement individuel non attenant	1	1	12
	Autre	2	2	21
Logement loué	Appartement	3	1	12
	Autre	4	2	21

Figure 3: Lignes du tableau de classement recoupé pour les personnes

Sexe	État matrimonial	Âge	N° de ligne	étiquette	Stratégie de combinaison		
					étape 1	étape 2	étape 3
Masculin	Jamais marié	0-4	1	1	12	123	
		5-9	2	2	21	213	
		10-14	3	3	32	321	
		15-19	4	1	12	123	1234
		20-24	5	2	21	213	2134
		25-44	6	3	34	342	3421
		45-64	7	4	43	432	4321
	Déjà marié	15-24	8	1	12		
		25-34	9	2	21		
		35-44	10	1	12	123	
		45-54	11	2	21	213	
		55-64	12	3	32	321	
		≥ 65	13	1			
Féminin	Jamais mariée	0-4	14	1	12	123	
		5-9	15	2	21	213	
		10-14	16	3	32	321	
		15-19	17	1	12	123	1234
		20-24	18	2	21	213	2134
		25-44	19	3	34	342	3421
		45-64	20	4	43	432	4321
	Déjà mariée	15-24	21	1	12		
		25-34	22	2	21		
		35-44	23	1	12	123	
		45-54	24	2	21	213	
		55-64	25	3	32	321	
		≥ 65	26	1			

Figure 4: Colonnes du tableau de classement recoupé pour les personnes

			Langue maternelle	No de col.	étiquette	Stratégie de combinaison étape 1	étape 2	étape 3
Membres d'une famille	Époux dans une famille époux-épouse	Sans enfant	Anglais (A)	1	1	13	132	132456
			Français (F)	2	2	23	231	231456
			Autre	3	3	31	312	312456
		Avec enfant(s)	A	4	4	46	465	465123
			F	5	5	56	564	564123
			Autre	6	6	64	645	645123
	Père ou mère dans une famille mono-parentale		A	7	1	13	132	
			F	8	2	23	231	
			Autre	9	3	31	312	
	Épouse dans une famille époux-épouse	Avec enfant(s) 0-5 ans	A	10	1	13	132	132456
			F	11	2	23	231	231456
			Autre	12	3	31	312	312456
		Aucun enfant 0-5 ans	A	13	4	46	465	465123
			F	14	5	56	564	564123
			Autre	15	6	64	645	645123
	Enfants dans les familles	0-14 ans	A	16	1	13	132	
			F	17	2	23	231	
			Autre	18	3	31	312	
		15 ans et plus	A	19	1	13	132	
			F	20	2	23	231	
			Autre	21	3	31	312	
Personnes hors familles	Personne 1		A	22	1	13	132	132456
			F	23	2	23	231	231456
			Autre	24	3	31	312	312456
	Autres membres du ménage		A	25	4	46	465	465123
			F	26	5	56	564	564123
			Autre	27	6	64	645	645123

Figure 5a: Différences population/échantillon pour les lignes du tableau de classement recoupé des ménages

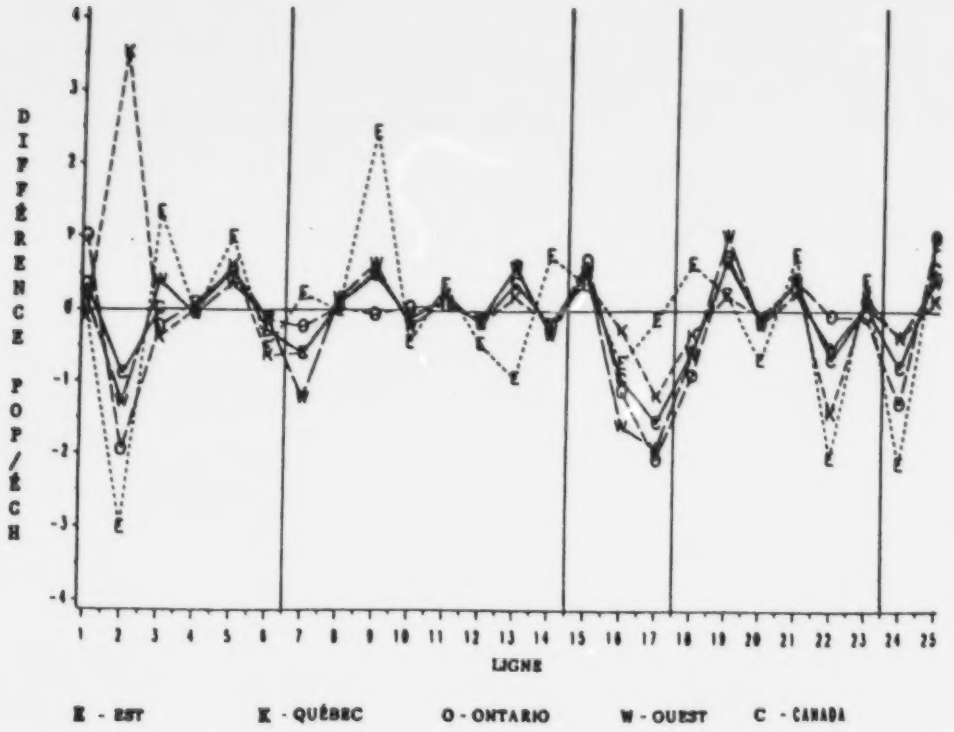


Figure 5b: Différences population/échantillon pour les colonnes du tableau de classement recoupé des ménages

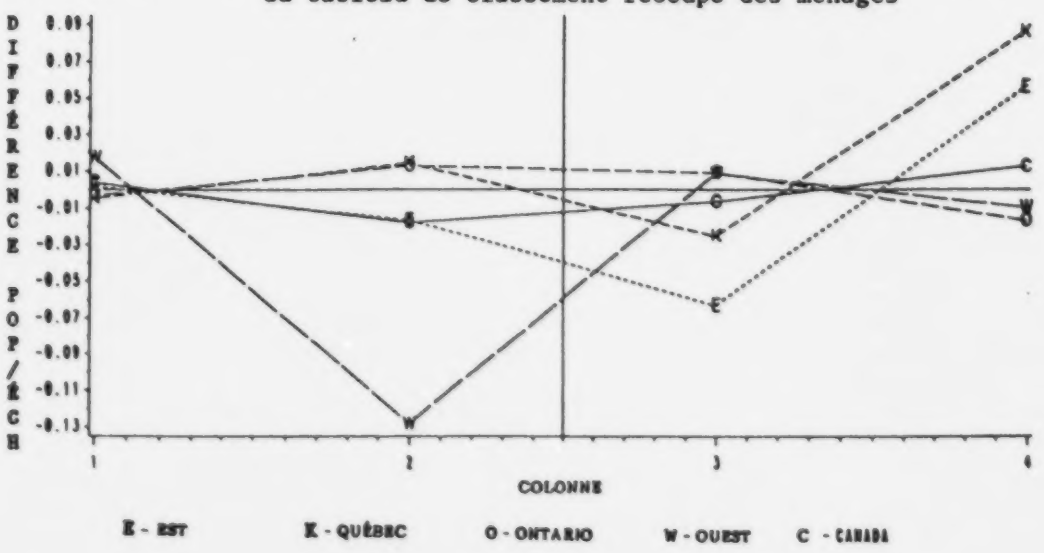


Figure 6a: Différences population/échantillon pour les lignes du tableau de classement recoupé des personnes

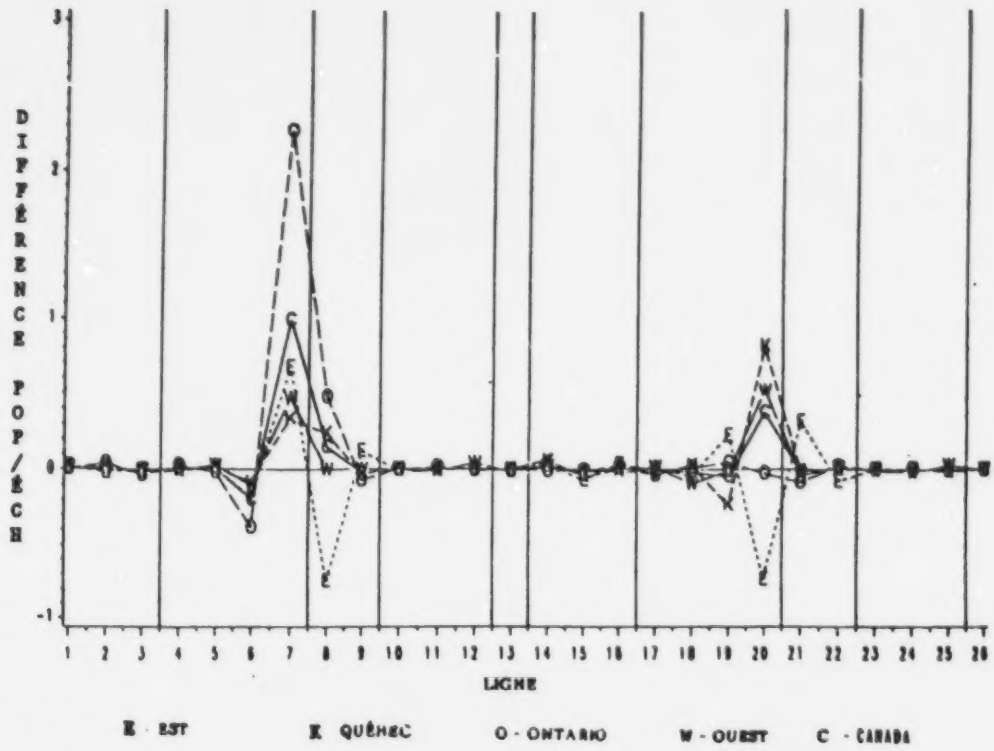
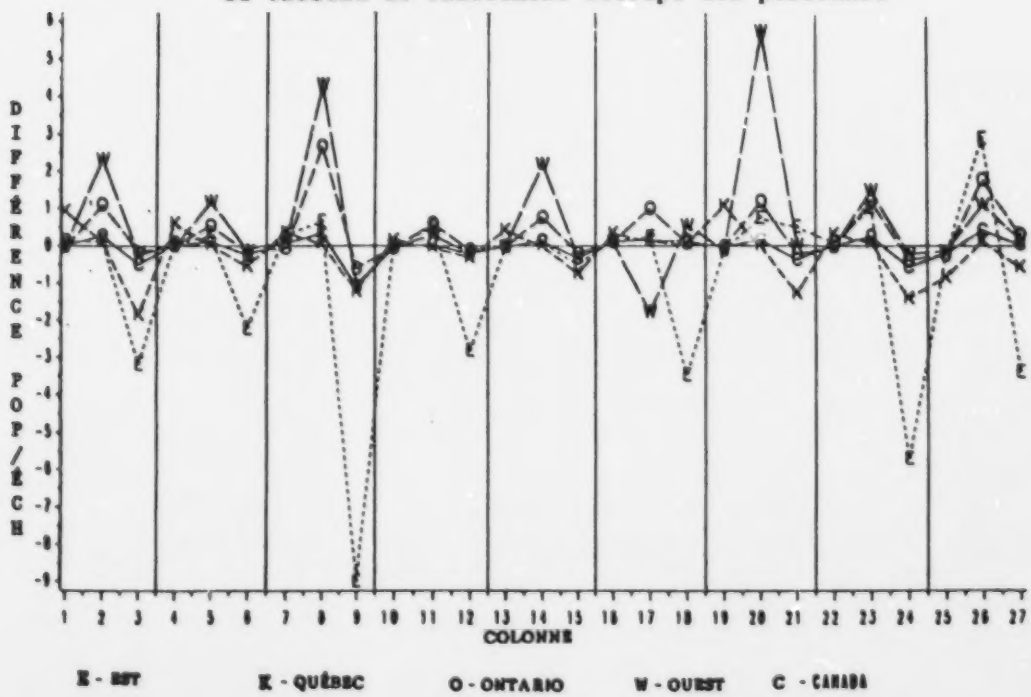


Figure 6b: Différences population/échantillon pour les colonnes du tableau de classement recoupé des personnes



ANNEXE II

FAMILLE DE RECENSEMENT

Époux et épouse (avec ou sans enfants jamais mariés, peu importe leur âge) ou un parent seul (peu importe son état matrimonial) avec un ou plusieurs enfants jamais mariés (quel que soit leur âge) vivant dans le même logement. Aux fins du recensement, les personnes vivant en union libre sont considérées comme "actuellement mariées", peu importe leur état matrimonial légal; elles figurent donc comme une famille époux-épouse dans la plupart des tableaux sur la famille de recensement.

Recensements: 1986, 1981, 1976, 1971, 1966, 1961

Remarques: La famille de recensement ainsi que son classement et ses variables connexes sont dérivés des réponses aux questions sur le lien avec la Personne 1, le sexe, la date de naissance et l'état matrimonial, et à partir de l'ordre dans lequel les membres du ménage paraissent sur le questionnaire.

FAMILLE ÉCONOMIQUE

Groupe de deux personnes ou plus qui vivent dans le même logement et qui sont apparentées par le sang, par alliance ou par adoption. Les personnes qui vivent en union libre sont considérées, aux fins du recensement, comme actuellement mariées, quel que soit leur état matrimonial légal; elles figurent donc comme des couples mariés dans les tableaux sur les familles économiques.

Recensements: 1986, 1981, 1976, 1971

Remarques: Pour former une famille économique, les membres d'un ménage doivent être apparentés par le sang, par alliance ou par adoption, tandis que pour former une famille de recensement, les membres d'une famille doivent avoir des liens conjugaux (époux-épouse) ou des liens de filiation (parent et enfant jamais marié). Par conséquent, la famille économique peut comprendre un plus grand nombre de personnes que la famille de recensement. Par exemple, une veuve vivant avec son fils et sa bru est considérée comme une personne hors famille, selon la définition de famille de recensement, mais fait partie de la famille économique incluant son fils et sa bru; deux familles apparentées, ou plus, vivant ensemble constituent également une famille économique (par exemple, un homme et son épouse vivant avec leur fils et leur bru); deux ou plusieurs frères ou soeurs vivant ensemble forment une famille économique, mais non une famille de recensement puisqu'ils ne répondent pas à la définition de cette dernière. Lorsque des personnes vivant ensemble sont apparentées, par exemple, deux frères, mais ne le sont pas à la Personne 1 du ménage (lorsque le lien avec la Personne 1 déclaré est, par exemple, chambreur, compagnon d'appartement ou employé), il est alors impossible de déterminer une famille économique à cause des contraintes du traitement des données.

La famille économique ainsi que son classement et ses variables connexes sont dérivés des réponses aux questions sur le lien avec la Personne 1, le sexe, la date de naissance et l'état matrimonial, et d'après l'ordre dans lequel les membres du ménage paraissent sur le questionnaire.

Le système général d'estimation pour les enquêtes économiques infra-annuelles[1]

Jean Dumais, Michel A. Hidiroglou[2]

Dans le cadre du projet de Remaniement des enquêtes-entreprises à Statistique Canada, plusieurs systèmes de traitement de l'information à usage général ont été développés. On a récemment entrepris de monter un "système général d'estimation". Ce document met en lumière les principaux aspects de ce système.

1 INTRODUCTION

Dans le cadre du projet de Remaniement des enquêtes-entreprises (Bilocq 1988, Hidiroglou 1988) et de plusieurs systèmes de traitement d'information à usage général (Bilocq 1988, Latouche 1988, Whitridge at al. 1988), il a été décidé d'élaborer un "système général d'estimation". Ce système devra être souple afin de pouvoir continuer de répondre aux besoins des enquêtes qui l'utiliseront. On a donc opté pour un développement modulaire.

1.1 Portée du Système Général d'Estimation:

L'équipe responsable du Système Général d'Estimation (SGE) doit fournir un ensemble de procédures utilisable par toutes les enquêtes économiques infra-annuelles. Cette précision est importante puisque les enquêtes économiques et les enquêtes sociales utilisent des méthodologies parfois incompatibles. De plus, les enquêtes annuelles et infra-annuelles n'utilisent habituellement pas le même type d'échantillonnage. En effet, ces dernières reposent souvent sur un plan d'échantillonnage à renouvellement partiel afin de réduire le fardeau de réponse.

Lors de sa mise en service, le SGE pourra être utilisé par les enquêtes mensuelles sur:

---

1) Présenté au colloque sur les méthodes et domaines d'application de la statistique du 57è congrès de l'Association Canadienne Française pour l'Avancement des Sciences, UQAM, mai 1989.

2) Division des méthodes d'enquête-entreprises, Statistique Canada, Ottawa, Onario, K1A 0T6.

- les Restaurants, traiteurs et tavernes,
- le Commerce de gros et de détail,
- les Stocks, livraisons et commandes, et
- l'Emploi, la rénumération et les heures de travail,

en plus des enquêtes trimestrielles sur:

- l'Organisation et les finances des entreprises, et
- le Transport routier des marchandises.

1.2 Plan d'échantillonnage de base:

On notera que les plans d'échantillonnage de ces enquêtes peuvent tous être vus comme un plan stratifié en grappes à un degré. Selon l'enquête, les grappes sont tantôt composées de toutes les unités d'échantillonnage formant un groupe de rotation, tantôt d'une seule unité d'échantillonnage. Il suffira donc de développer un module d'estimation pour un seul plan d'échantillonnage afin de répondre aux besoins de toutes les enquêtes énumérées ci-haut.

1.3 Types d'estimation nécessaires:

Le SGE devra permettre à l'usager de produire des estimations de totaux, de moyennes, de quotients, ou de tendances. Chaque estimation sera accompagnée d'une estimation de la variabilité (coefficient de variation ou éventuellement erreur relative).

De plus, on devra être en mesure de fournir des estimations par domaines, selon les besoins de l'utilisateur.

2 LE SYSTÈME GÉNÉRAL D'ESTIMATION

2.1 Interfaces:

La notion même de système généralisé contient celle de paramétrage. On préférera programmer en fonction de quelques plans d'échantillonnage courants et que l'usager précise les particularités de son enquête, à la programmation de tous les plans d'échantillonnage en autant de systèmes distincts.

L'usager identifiera dans un premier temps le plan d'échantillonnage régissant son enquête: il s'agira de fournir au SGE la hiérarchie des strates, unités primaires d'échantillonnage et unités secondaires d'échantillonnage.

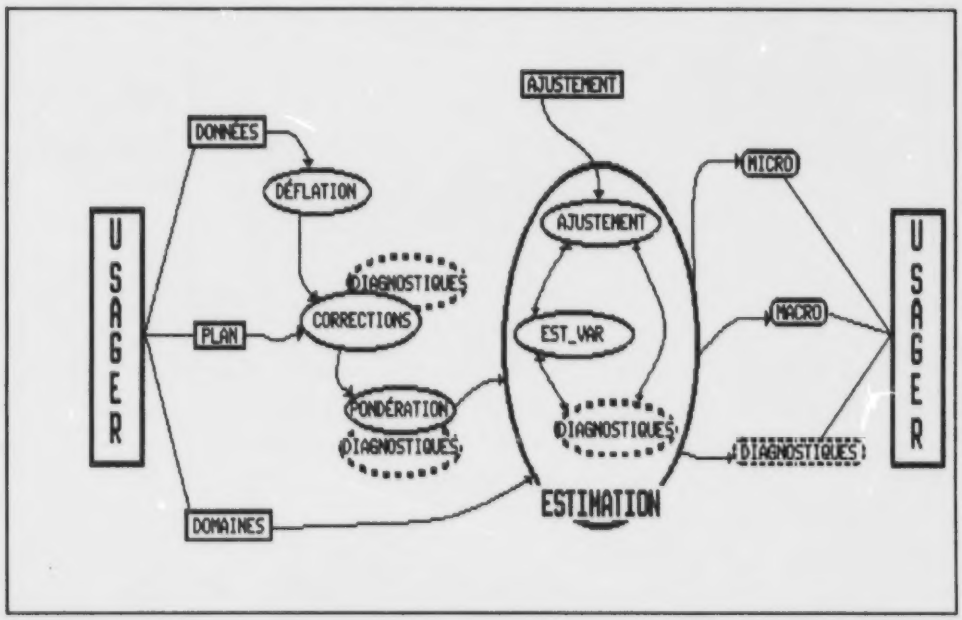
Un deuxième ensemble de paramètres permettra de faire l'identification des domaines d'estimation, alors que les variables d'intérêt pour lesquelles on désire ces estimations seront précisées dans un troisième ensemble de paramètres.

Enfin, le choix du type d'estimateur est le dernier paramètre que l'utilisateur fournira au SGE.

Ces quatre ensembles de paramètres sont essentiels et devront être founis au SGE; l'absence de l'un ou l'autre de ces ensembles signifiera que l'usager est satisfait des valeurs par défaut.

Une fois ces paramètres fournis, les différents modules du SGE entreront en action suivant la commande passée par l'usager.

2.2 Modules principaux:



Le SGE sera construit autour de quatre modules principaux: déflation, correction, pondération et estimation.

Les deux premiers modules sont des modules de correction. Le premier, le module de déflation, sert à modifier la valeur prise par une des variables d'enquête pour une unité d'échantillonnage donnée. C'est une correction du type "winsorisation". Le deuxième module, celui de correction, sert à modifier le poids d'estimation associé à une unité d'échantillonnage. Contrairement à la déflation, ce type de modification affecte toutes les variables d'enquête d'une unité. Ces corrections ne sont valables que pour la période en cours de traitement et ne sont pas conservées d'une occasion à l'autre.

Ces modules peuvent être considérés optionnels (par opposition aux paramètres de la section précédente) dans la mesure où les valeurs par défaut sont de "ne rien corriger".

Le module de pondération est celui où les poids d'estimation seront calculés en accord avec le type d'estimateur choisi. Enfin, les estimations par domaines et l'estimation de l'erreur d'échantillonnage seront assemblés par le module d'estimation.

Un cinquième module, optionnel lui aussi, verra aux ajustements post-estimation.

3 ESTIMATEURS PROPOSÉS

Le groupe de travail responsable du développement du SGE proposera aux utilisateurs quatre estimateurs du total lors de la mise en service initiale du système.

Dans ce qui suit, Y représente la variable d'enquête, h la strate, i l'u.p.e., et t la période pour laquelle on estime. Les tailles de population et d'échantillon de la strate h sont respectivement données par P_h et p_h pour le nombre de «grappes» (ou groupes de rotation), et N_h et n_h pour le nombre d'u.s.e..

Le plan d'échantillonnage de base suppose que l'on a tiré un échantillon aléatoire simple de p_h «grappes» parmi les P_h «grappes» de la strate. On suppose de plus qu'aucune correction (voir 2.2) n'a été demandée. Si aucun paramètre de commande d'estimateur n'est passé au système, l'usager choisit par défaut l'estimateur de Mickey.

3.1 Estimateur par valeurs dilatées:

Le premier estimateur proposé est l'estimateur par valeurs dilatées:

$$\hat{Y}_{ed} = \sum_h \frac{P_h}{p_h} \sum_i y_{hi}$$

3.2 Estimateur par le quotient (séparé):

Contrairement à l'estimateur par valeurs dilatées, l'estimateur par le quotient tient compte du nombre d'u.s.e. observées:

$$\hat{Y}_{qs} = \sum_h \frac{N_h}{n_h} \sum_i y_{hi}$$

3.3 Estimateur par le quotient de Mickey:

On trouve dans la littérature (Rao 1969, Rao et Rao 1971, Cochran 1977) un certain nombre d'estimateurs du type quotient dits "sans biais". L'estimateur de Mickey (1959) est l'un de ceux-là. On peut voir cet estimateur comme une version "après jackknife" de l'estimateur par le quotient:

$$\hat{Y}_M = \sum_h \sum_i \{(N_h - n_h(P_h - p_h + 1))d_{hi} + (P_h - p_h + 1)\} y_{hi}$$

avec $$d_{hi} = \frac{1}{p_h}\left\{\left[\sum_{k=1}^{p_h}\left(\frac{1}{n_h - N_{hk}}\right)\right] - \left(\frac{1}{n_h - N_{hi}}\right)\right\}$$

3.4 Estimateur de Mickey composite:

On a de plus proposé une amélioration à l'estimateur de Mickey: on construit un estimateur composite utilisant les estimations de Mickey pour les périodes précédente et courante:

$$\hat{Y}_C^{(t)} = \alpha \hat{Y}_M^{(t)} + (1 - \alpha)\{\hat{Y}_M^{(t-1)} + \hat{D}^{(t-1,t)}\}$$

où $\hat{D}^{(t-1,t)}$ estime l'évolution dans le temps.

3.5 Estimation de la variance:

On a choisi d'estimer la variance des estimateurs en utilisant la méthode du "jackknife" . On sait que dans les cas simples, l'estimation

par "jackknife" est identique à celle obtenue par les méthodes analytiques (Wolter 1985). De plus, le "jackknife" est une façon commode d'obtenir une estimation de la variance de l'estimateur de Mickey.

Si l'estimation de la variance par "jackknife" dans une strate est (en omettant l'indice h):

$$JK(\hat{Y}) = \frac{1}{p-1}\sum_{j=1}^{p}(\hat{Y}^{(j)} - \bar{\hat{Y}})^2$$

avec

$$\bar{\hat{Y}} = \frac{1}{p}\sum_{j=1}^{p}\hat{Y}^{(j)}$$

et

$$\hat{Y}^{(j)} = \sum_{i} w_i^{(j)} y_i$$

alors les poids d'estimation sont:

$$w_i^{(j)} = \begin{cases} \dfrac{P}{p-1} & \text{si} \quad i \neq j \\ 0 & \text{si} \quad i = j \end{cases}$$

$$w_i^{(j)} = \begin{cases} \dfrac{N}{n-N_j} & \text{si} \quad i \neq j \\ 0 & \text{si} \quad i = j \end{cases}$$

$$w_i^{(j)} = \begin{cases} (N-(n-N_j)(P-p+2))d_i^{(j)} + (P-p+2) & \text{si} \quad i \neq j \\ 0 & \text{si} \quad i = j \end{cases}$$

où

$$d_i^{(j)} = \frac{1}{p-1}\left\{\left[\sum_{k=1}^{p}\left(\frac{1}{n-N_k-N_i}\right)\right] - \left(\frac{1}{n-2N_j}\right) - \left(\frac{1}{n-N_j-N_i}\right)\right\}$$

pour les estimateurs par valeurs dilatées, par le quotient, et de Mickey respectivement (Dumais, Carpenter 1988).

3.6 Un exemple:

Considérons l'exemple suivant. On dispose d'une population de N=20 unités découpée en P=5 «grappes» tel qu'illustré dans le diagramme ci-dessous; de cette population, on a choisi un échantillon aléatoire simple de p=3 «grappes». On donne les poids d'estimation associés à chaque «grappe» pour chacun des trois estimateurs principaux:

	grappes					
Tailles N_i:	0	3	7	4	6	
$\hat{Y}_M$:	1.40	1.55	2.19	0	0	$w_i=((N-n(P-p+1))d_i + (P-p+1))y_i$
$\hat{Y}_{vd}$:	5/3	5/3	5/3	0	0	$w_i=P/p$
$\hat{Y}_{qs}$:	20/10	20/10	20/10	0	0	$w_i=N/n$

grappes échantillonnées

4 VALIDATIONS

4.1 Comparaison des estimateurs:

Les différences théoriques entre les quatre estimateurs ont évidemment des répercussions sur l'utilisation. Le tableau ci-dessous résume les avantages et désavantages des différents estimateurs:

Estimateur	Pour	Contre
Valeurs dilatées	simple sans biais	inefficace
Quotient séparé	simple efficace	biaisé
Estimateur de Mickey	sans biais efficace	plus de calculs poids négatifs possibles
Mickey composite	sans biais efficacité maximum	maximum de calculs

4.2 Simulation:

Choudhry et Gossen (1988) ont étudié le biais et l'efficacité relatifs des trois premiers estimateurs proposés par rapport à l'estimateur par valeurs dilatées $\hat{Y}_{vd}$.

Ils ont utilisé les données de l'Enquête mensuelle sur le commerce de détail: la "population" a été coupée en 32 strates de P=16 «grappes», et ils ont pigé tous les échantillons de p=4 «grappes».

Les résultats présentés ci-dessous sont ceux de quatre groupes de simulations. Dans un premier temps, on a constitué des «grappes» comptant environ le même nombre d'u.s.e. («grappes» équilibrées); puis on a changé certaines u.s.e. de «grappes» créant ainsi des «grappes» déséquilibrées. Dans chacun des cas, on a construit deux répliques: l'une où chaque u.s.e. fournit une réponse positive (bonne corrélation entre Y et N_i), l'autre où un certain nombre d'u.s.e. ne contribue pas à l'estimation (mauvaise corrélation).

		taille des «grappes»	
	groupe	équilibrées	déséquilibrées
corrélation	bonne	1	3
entre Y_i et N_i	mauvaise	2	4

Dans le tableau qui résume les simulations (ci-dessous), les biais (B) et efficacité (E) relatifs sont exprimés en pourcentage; par exemple, pour le groupe 1, la variance de l'estimateur de Mickey était 2 fois plus petite que celle de l'estimateur par valeurs dilatées.

groupe	$CV(N_i)$	$\rho_{N,Y}$		EVD	EQS	EM
1	0.09	0.81	B	0	-0.14	0
			E	100	212	207
2	0.06	0.19	B	0	-0.06	0
			E	100	103	103
3	0.34	0.61	B	0	-0.56	0
			E	100	155	147
4	0.36	0.18	B	0	1.38	0
			E	100	92	95

5 A VENIR

Au cours des mois à venir, voire des quelques prochaines années, le Système Général d'Estimation bénéficiera de plusieurs ajouts importants.

On ajoutera un module de post-stratification pour permettre un ajustement plus fin des estimations.

On incorporera au plan de sondage actuel les plans avec probabilité proportionnelle à la taille avec remise, et les plans à plusieurs degrés.

On ajoutera à la batterie d'estimateurs disponibles l'estimateur par la régression généralisée (Särndal et coll. 1989).

BIBLIOGRAPHIE

Bilocq, F., "La fonction générale de collecte et de saisie", *Recueil des textes des présentations du colloque sur les méthodes et domaines d'application de la statistique*, ACFAS 1988, Bureau de la statistique du Québec, Québec, 1988.

Choudhry, G.H., *Generalized Estimation and Variance Estimation System for Sub-annual Surveys*, Statistique Canada, document interne, avril 1988.

Choudhry, G.H., Gossen, M., Statistique Canada, document interne sans titre, 1988.

Cochran, W.G., *Sampling Techniques*, 3rd edition, John Wiley, New York, 1977.

Dumais, J., *Generalized Estimation System Interfaces*, Statistique Canada, DMEE, décembre 1988.

Dumais, J., Carpenter, R., *Methodology of the Generalized Estimation System for Sub-annual Business Surveys*, Statistique Canada, DMEE, septembre 1988.

Hidiroglou, M.A., "La méthodologie pour les enquêtes mensuelles de commerce de gros et de détail", *Recueil des textes des présentations*

du colloque sur les méthodes et domaines d'application de la statistique, ACFAS 1988, Bureau de la statistique du Québec, Québec, 1988.

Hidiroglou, M.A., Srinath, K.P., *A General Strategy for Selection, Rotation and Maintenance of Samples for Business Surveys*, Statistique Canada, document interne, décembre 1987.

Laniel, N., "Variances for a rotating sample from a changing population", in Proceedings of the Survey Research Section, American Statistical Association, 1987.

Latouche, M., "Détermination, allocation et sélection de l'échantillon", *Recueil des textes des présentations du colloque sur les méthodes et domaines d'application de la statistique*, ACFAS 1988, Bureau de la statistique du Québec, Québec, 1988.

Mickey, M.R., "Some finite population unbiased ratio and regression estimators", *J.A.S.A.*, vol. 54, 1959.

Rao, J.N.K., "The precision of Mickey's unbiased ratio estimator", Biometrika, vol.54, 1967.

Rao, J.N.K., "Ratio and regression estimators", in New Developments in Survey Sampling, Johnson, N.L., Smith, H., ed. Wiley-Interscience, 1969.

Rao, P.S.R.S., Rao, J.N.K., "Small sample results for ratio estimators", Biometrika, vol.58, 1971.

Särndal, C.E., Swensson, B., Wretman, J.H., "The Weighted residual technique for estimating the variance of the general regression estimator", à paraitre, *Biometrika*, 1989.

Srinath, K.P., *Generalized Estimation Procedure for Continuous Business Surveys*, Statistique Canada, séminaires DMEE, 1988.

Tambay, J.L., *Extreme Weights with Mickey and Quenouille Estimators under some Restrictions*, notes manuscrites, juillet 1988.

Whitridge, P., Macmillan, J., Kovar, J., "Système généralisé de vérification et d'imputation pour les enquêtes économiques à Statistique Canada", *Recueil des textes des présentations du colloque sur les méthodes et domaines d'application de la statistique*, ACFAS 1988, Bureau de la statistique du Québec, Québec, 1988.

Wolter, K.M., *Introduction to Variance Estimation*, Springer-Verlag, New York, 1985.

De la Robustesse dans l'Estimation de la Moyenne d'une Population Finie: une Généralisation de la Méthode du Quotient.

Jean-Philippe GWET
(Université Laval)

1. RÉSUMÉ.

Sur un exemple concret, nous montrons dans cet éxposé que l'estimateur obtenu par la méthode du quotient est très sensible à la présence des valeurs aberrantes dans un échantillon. Pour remédier à ce problème, certaines considérations modèlistes nous amèneront à une première généralisation de la méthode. Cette généralisation qui se fait dans le cas d'une population finie est analogue à l'approche du problème de la robustesse par le M-estimateur dans le cas de l'estimation paramétrique sur des populations infinies. Il sera montré par la suite que le M-estimateur reste sensible aux valeurs aberrantes d'un type particulier. Cette observation nous conduira à une deuxième généralisation de l'estimateur du quotient qui est l'estimateur de Mallows et qui répond mieux au problème qui nous préocuppe selon le critère d'efficacité. L'approche ici se fera par le plan d'échantillonnage et les résultats d'une étude expérimentale seront présentés à la fin.

2. INTRODUCTION.

On considère une population U composée de N unités de sondages. La lettre Y désigne la variable d'intérêt et $y_1, \ldots, y_N$ les valeurs qu'elle prend dans chaque unité de sondage. La lettre X désigne la variable auxilliaire et les valeurs qu'elle prend sur les unités de sondages seront notées $x_1, \ldots, x_N$.
Il s'agit de faire une estimation du paramètre $\overline{Y}$ défini par:

$$\overline{Y} = \frac{1}{N}\sum_{i=1}^{N} y_i$$

La lettre S désigne l'ensemble de tous les échantillons de taille n pouvant être tirés de la population U; tandis que s désignera un élément de cet ensemble. Nous supposerons tout au long de cet exposé que les échantillons s sont aléatoires simples, que les valeurs x_i $(i = 1, \ldots, N)$ sont connues avant l'enquête et que les valeurs aberrantes se trouvant dans un échantillon ne résultent pas des erreurs de transcription.

3. ESTIMATION PAR LE QUOTIENT.

Le paramètre $\overline{Y}$ que nous voulons estimer peut s'écrire sous la forme suivante:

$$\overline{Y} = \left(\frac{\overline{Y}}{\overline{X}}\right)\overline{X}$$

$$= R\overline{X} \quad \text{où } (\mathrm{R} = \text{quotient} = \frac{\overline{Y}}{\overline{X}})$$

Puisque le paramètre $\overline{X}$ est connu, il suffirait d'avoir une estimation de R pour obtenir une estimation de $\overline{Y}$.

La méthode du quotient consiste donc à ramener le problème de l'estimation du paramètre $\overline{Y}$, en un problème d'estimation du quotient des deux paramètres $\overline{Y}$ et $\overline{X}$. Une façon naturelle d'estimer le quotient R à partir d'un échantillon s est de poser:

$$\hat{R} = \frac{\overline{y}}{\overline{x}} \quad \text{où} \begin{cases} \overline{y} = (1/n) \sum_{i \in s} y_i \\ \overline{x} = (1/n) \sum_{i \in s} x_i \end{cases}$$

Remarque.
L'intérêt de la méthode du quotient peut s'observer de la manière suivante: si nous notons l'estimateur quotient du paramètre $\overline{Y}$ par $\hat{\overline{Y}}_q$ alors

$$\begin{aligned} \hat{\overline{Y}}_q &= \hat{R}\,\overline{X} \\ &= \left(\frac{\overline{X}}{\overline{x}}\right)\overline{y} \end{aligned}$$

Il est alors nécessaire de noter que $\overline{x}$ et $\overline{y}$ sont des estimateurs de $\overline{X}$ et $\overline{Y}$ respectivement. Si l'échantillon s conduit à une sous-estimation (resp. surestimation) de $\overline{X}$ et que le même phénomène s'observe au niveau de la variable Y (ce qui suppose une corrélation positive entre X et Y) alors la multiplication de $\overline{y}$ par le facteur $\overline{X}/\overline{x}$ sera un réajustement à la hausse (resp. baisse) de la valeur de $\overline{y}$. Cette méthode permet en effet d'améliorer la précision de l'estimation.

Toutefois, comme on le remarquera à l'exemple suivant, l'estimateur $\hat{R}$ peut avoir une variance anormalement élévée si l'échantillon s qui est à la base de l'estimation contient des valeurs aberrantes.

Exemple 1:
On considère une population de 34 villages de la subdivision de Lucknow en Inde qu'on va nommer la population I. La variable d'intérêt Y représente la surface de blé en 1936 dans un village et la variable auxilliaire X, la surface totale cultivée en 1931. On peut remarquer sur la figure 1 la présence d'une observation P qui est substantiellement éloignée des autres. En supprimant cette observation, on obtient la population II. Ces deux populations sont décrites dans la tableau 1 ci-dessous.

TABLE 1 : DÉSCRIPTION DES POPULATIONS I ET II						
POP.	SOURCE	$\overline{Y}$	$\overline{X}$	N	$C_x^{(*)}$	$\rho^{(**)}(X,Y)$
I	S. & S.(1970, p. 185)	218.41	765.35	34	0.62	0.83
II	S. & S.(1970, p. 185)	204.94	762.94	33	0.63	0.92

(*)coefficient de variation de X
(**)coefficient de corrélation linéaire entre X et Y.

21 échantillons aléatoires simples ont été tirés de la population 1 à partir desquels on a fait 21 estimations de $\overline{Y}$ par la méthode du quotient. Les résultats sont présentés sur la figure 2 ci-dessous.

fig. 1: graphique des observations de la population 1

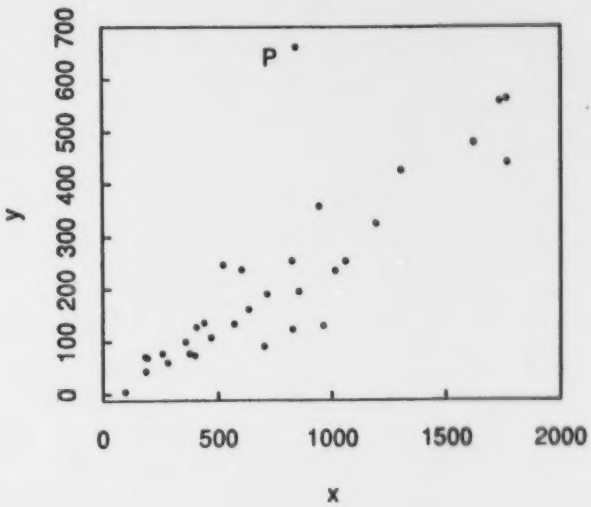
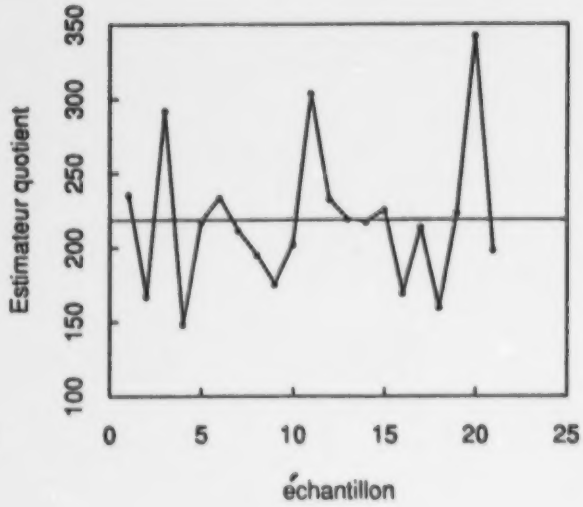
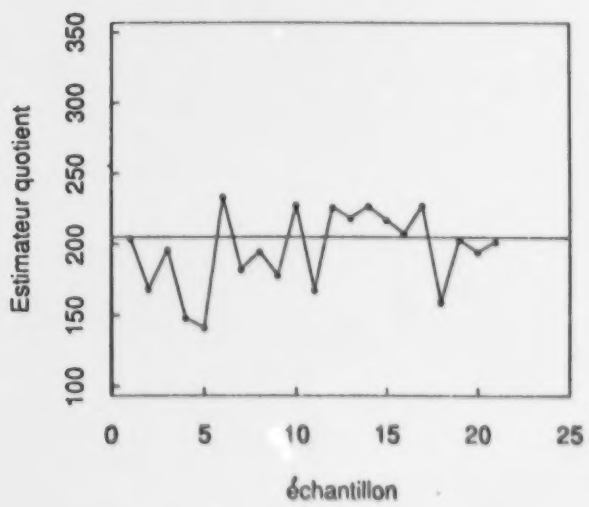


fig. 2: 21 estimations faites par la méthode du quotient. (pop. 1)



On peut remarquer sur la figure 2 que les trois échantillons 3, 11 et 20 ont donnés une surestimation importante de la vraie valeur du paramètre $\overline{Y}$ qui est représentée par la ligne horizontale. En faisant 21 estimations de $\overline{Y}$ à l'aide de 21 échantillons tirés de la population 2, on obtient les résultats de la figure 3.

fig. 3: 21 estimations faites par la méthode du quotient à partir d'échantillons tirés de la population 2.



4. ESTIMATION ROBUSTE. LE M-ESTIMATEUR.

Dans le cadre de l'approche par le modèle, BREWER(1963) et ROYALL(1970b) ont établi le résultat suivant:

Théorème.

Soit le modèle (M) suivant:

$$\begin{cases} y_i = \beta x_i + \epsilon_i \\ E_m(\epsilon_i) = 0 \\ V_m(\epsilon_i) = \sigma^2 x_i \\ E_m(\epsilon_i x_i) = 0 \end{cases}$$

Alors, sous le modèle (M), l'estimateur du quotient est le meilleur estimateur linéaire sans biais du paramètre $\overline{Y}$.

L'intérêt pratique de ce résultat réside dans le fait que si le graphique des observations (x_i, y_i) $(i = 1, \ldots, n)$ indique que celles-ci peuvent être décrites par un modèle comparable à (M), alors l'utilisation de l'estimateur quotient serait commode. Pour la suite, le modèle (M) sera le modèle de référence.

Afin de résoudre le problème posé par la présence des valeurs aberrantes dans l'échantillon, on se fixe comme objectif de trouver un estimateur T_n qui remplit les trois conditions suivantes:

1- l'estimateur devra avoir une précision comparable à celle de l'estimateur quotient sous le modèle implicite (M);

2- une légère perturbation des données ne pourra affecter que très peu la distribution de T_n;

3- une grosse perturbation des données n'affectera pas la validité de l'estimateur.

En se référant au modèle (M), rappelons que l'estimateur quotient de $\overline{Y}$ est donné par

$$\widehat{\overline{Y}}_q = \widehat{R}\,\overline{X}$$

où $\widehat{R}$ est solution de l'équation suivante:

$$\sum_{i \in s} \sqrt{x_i} \left(\frac{y_i - \beta x_i}{\sigma \sqrt{x_i}} \right) = 0$$

Cette équation peut aussi s'écrire sous la forme

$$\sum_{i \in s} \sqrt{x_i}\, \psi \left(\frac{y_i - \beta x_i}{\sigma \sqrt{x_i}} \right) = 0 \quad \text{où } \psi(r) = r\,,\ \forall r \in \mathbb{R} \tag{1}$$

Introduisons le paramètre β_M implicitement défini comme solution de l'équation

$$\sum_{i=1}^{N} \sqrt{x_i}\, \psi \left(\frac{y_i - \beta x_i}{\sigma \sqrt{x_i}} \right) = 0 \tag{2}$$

Remarques.

1) β_M est un paramètre qui est directement relié à la population U et que nous nous proposons d'estimer. Nous appelerons M-estimateur de β_M basé sur la fonction ψ toute solution $\widehat{\beta}_M$ de l'équation (1).

2) Si la fonction ψ est la fonction identité, alors

$$\beta_M = \frac{\overline{Y}}{\overline{X}} = R$$
$$\hat{\beta}_M = \frac{\overline{y}}{\overline{x}} = \hat{R}$$

3) Les enquêtes étant faites de façon répétitives, on a généralement une assez bonne idée de la variabilité des données. Nous supposerons donc que la constante σ est connue. De plus nous verrons plus loin avec le choix de la fonction ψ qu'on peut supposer sans nuire à la généralité que $\sigma = 1$.

Nous nous proposons présentement d'étudier l'estimateur $\hat{\beta}_M$ pour une fonction ψ quelconque.
Il est possible d'établir que pour une taille d'échantillon suffisamment élévée, $\hat{\beta}_M$ et β_M sont liés par la relation suivante:

$$\hat{\beta}_M \simeq \beta_M + \frac{(1/n)\sum_{i\in s} \sqrt{x_i}\,\psi\left(\frac{y_i - \beta_M x_i}{\sqrt{x_i}}\right)}{(1/N)\sum_{i=1}^{N} x_i \psi'\left(\frac{y_i - \beta_M x_i}{\sqrt{x_i}}\right)} \tag{3}$$

• le M-estimateur de $\overline{Y}$ est alors donné par:

$$\hat{\overline{Y}}_M = \hat{\beta}_M \overline{X}$$

• A l'aide de la relation (3) ci-dessus, on établit que l'estimateur $\hat{\overline{Y}}_M$ est biaisé et que son biais asymptotique est donné par:

$$B(\hat{\overline{Y}}_M) = \beta_M \overline{X} - \overline{Y} \tag{4}$$

Remarque. *Si les valeurs aberrantes sont distribuées de façon symmétrique, ou bien alors si les données sont peu perturbées, on peut s'attendre à un biais relativement faible.*

• *L'erreur quadratique moyenne du M-estimateur de $\overline{Y}$ est donnée par:*

$$EQM(\hat{\overline{Y}}_M) = [B(\hat{\overline{Y}}_M)]^2 + \left(\frac{\overline{X}}{\overline{\mu}}\right)^2 \frac{1-f}{n(N-1)} \sum_{i=1}^{N} x_i \left[\psi\left(\frac{y_i - \beta_M x_i}{\sqrt{x_i}}\right)\right]^2$$

$$\text{où} \begin{cases} \overline{\mu} = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} \mu_i \\ \mu_i = x_i \psi'\left(\dfrac{y_i - \beta_M x_i}{\sqrt{x_i}}\right) \end{cases}$$

On peut remarquer que si ψ est la fonction indentité, alors $\mu_i = x_i$ et on a

$$EQM(\hat{\overline{Y}}_M) = \frac{1-f}{n(N-1)} \sum_{i=1}^{N} \left(y_i - \frac{\overline{Y}}{\overline{X}} x_i\right)^2$$

on retrouve ainsi la formule classique de la variance de l'estimateur quotient.

• **Un estimateur de l'erreur quadratique moyenne est donné par:**

$$eqm(\hat{\bar{Y}}_M) = [b(\hat{\bar{Y}}_M)]^2 + \left(\frac{\overline{X}}{\overline{\mu}_s}\right)^2 \frac{1-f}{n(n-1)} \sum_{i \in s} x_i \left[\psi\left(\frac{y_i - \hat{\beta}_M x_i}{\sqrt{x_i}}\right)\right]^2$$

$$\text{où} \begin{cases} b(\hat{\bar{Y}}_M) = \hat{\beta}_M \bar{x} - \bar{y} & \textit{(estimation du biais)} \\ \overline{\mu}_s = (1/n) \sum_{i \in s} \mu_i \end{cases}$$

Remarques.

1) Lorsque la taille de l'échantillon est faible, il devient assez difficile d'obtenir une bonne estimation du biais. Dans ce cas, l'estimation de la variance constitue généralement une meilleure estimation de l'erreur quadratique moyenne.

2) Si la fonction ψ est la fonction identité, alors

$$eqm(\hat{\bar{Y}}_M) = \left(\frac{\overline{X}}{\bar{x}}\right)^2 \frac{1-f}{n(n-1)} \sum_{i \in s} \left(y_i - \frac{\bar{y}}{\bar{x}} x_i\right)^2$$

On retrouve un estimateur de la classe des estimateurs de WU(1982) avec $g = 2$.

• CHOIX DE LA FONCTION ψ.

Pour des raisons d'efficacité, on choisi généralement une fonction ψ impaire telle que:

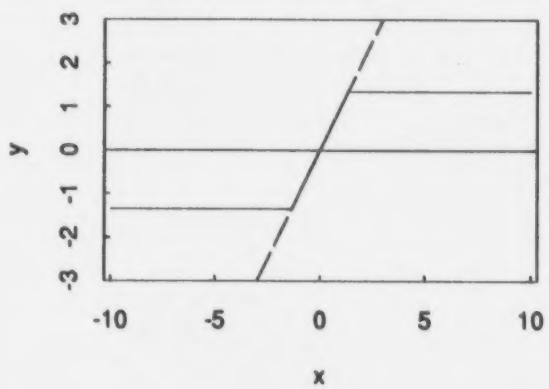
$$\lim_{t \to 0} \frac{\psi(t)}{t} = 0$$

et pour des raisons de robustesse, la fonction ψ devra être bornée.

Exemples.

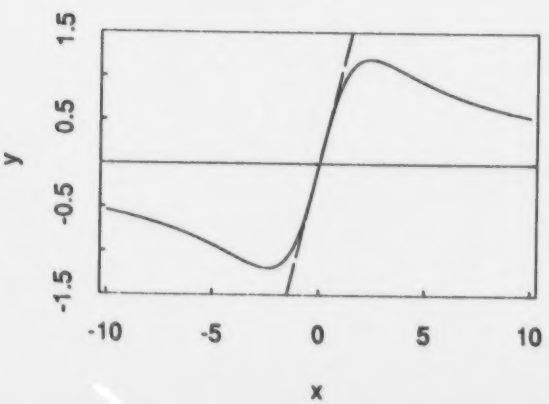
1) (Huber, 1964)

$$\psi_h(x) = \begin{cases} -h \text{ si } x \leq -h \\ x \text{ si } -h < x < h \\ h \text{ si } x \geq h \end{cases}$$



2) Fonction de Cauchy

$$\psi_c(x) = \frac{x}{1 + \left(\frac{x}{c}\right)^2}$$



• CALCUL DE L'ESTIMATEUR $\widehat{\beta}_M$.

L'estimateur $\widehat{\beta}_M$ est solution de l'équation (1) qui peut aussi s'écrire sous la forme suivante:

$$\sum_{i\in s}\omega_i x_i(y_i-\xi x_i)=0$$

$$\text{où}\quad \omega_i=\begin{cases}\dfrac{1}{\sqrt{x_i}\,(y_i-\xi x_i)}\psi\left(\dfrac{y_i-\xi x_i}{\sqrt{x_i}}\right) & \text{si } y_i-\xi x_i\neq 0\\ \dfrac{1}{x_i} & \text{sinon}\end{cases}$$

Etant donné que le vecteur de poids $\omega=(\omega_1,\ldots,\omega_n)$ dépend de l'inconnue ξ, il n'est pas possible de trouver une solution explicite de cette équation. Nous allons pour cela utiliser la méthode itérative de Beaton & Tukey(1974). Elle consiste à calculer à chaque itération le vecteur de poids ω jusqu'à la convergence.
Formellement nous avons:

$$\widehat{\beta}_M^{(t+1)}=\widehat{\beta}_M^{(t)}+\frac{\sum_{i\in s}\omega_i^{(t)}x_i(y_i-\widehat{\beta}_M^{(t)}x_i)}{\sum_{i\in s}\omega_i^{(t)}x_i^2}$$

$$\text{où}\quad \omega_i=\begin{cases}\dfrac{1}{\sqrt{x_i}\,(y_i-\widehat{\beta}_M^{(t)}x_i)}\psi\left(\dfrac{y_i-\widehat{\beta}_M^{(t)}x_i}{\sqrt{x_i}}\right) & \text{si } y_i\neq\widehat{\beta}_M^{(t)}x_i\\ \dfrac{1}{x_i} & \text{sinon}\end{cases}$$

si la fonction ψ est croissante (e.g. celle de Huber) alors le point de convergence est unique et on peut prendre comme valeur initiale

$$\widehat{\beta}_M^{(0)}=\frac{\overline{y}}{\overline{x}}=\widehat{R}$$

Nous verrons dans l'étude expérimentale de la section 6, que l'utilisation du M-estimateur de $\overline{Y}$ peut se traduire par un gain considérable en efficacité.

Cependant l'exemple suivant nous permet de voir que le M-estimateur est aussi sensible à des valeurs aberrantes d'un type particulier: celles pour lesquelles la valeur de X est assez élévée.

Exemple 2.

Considérons l'ensemble de 20 données décrite par la figure 4 ci-dessous. Nous remarquons sur ce graphique la présence d'une observation P qui appartient au nuage de points. La figure 5 représente l'efficacité du M-estimateur par rapport à l'estimateur quotient quand on fait augmenter l'abscisse $x(p)$ de P en lui attribuant les valeurs 200, 300, 400, 500, 600, 700, 800, 900 et 1000. Pour chaque valeur de $x(p)$ tous les échantillons de taille $n=3$ ont été tirés pour caculer $EQM(\widehat{\overline{Y}}_q)$ et $EQM(\widehat{\overline{Y}}_M)$. L'estimateur $\widehat{\overline{Y}}_M$ est basé sur la fonction ψ de Huber avec $h=1.345\hat{\sigma}$

$$\text{où}\quad \hat{\sigma}=1.48\text{méd}|t_i-\text{méd}(t_i)|\quad\text{et } t_i=\frac{y_i-\widehat{R}x_i}{\sqrt{x_i}}$$

fig. 4: graphique des observations de la population.

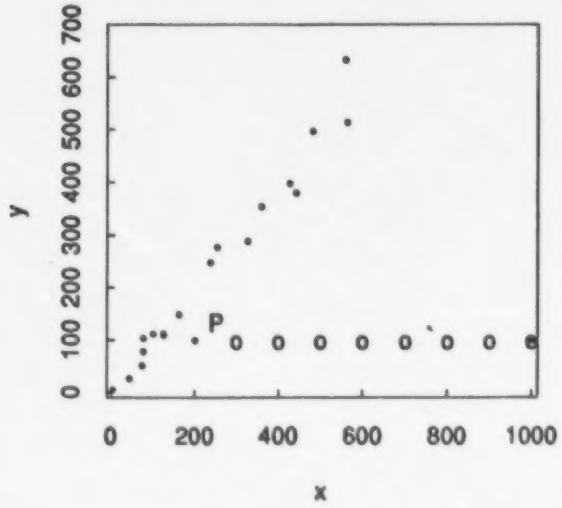
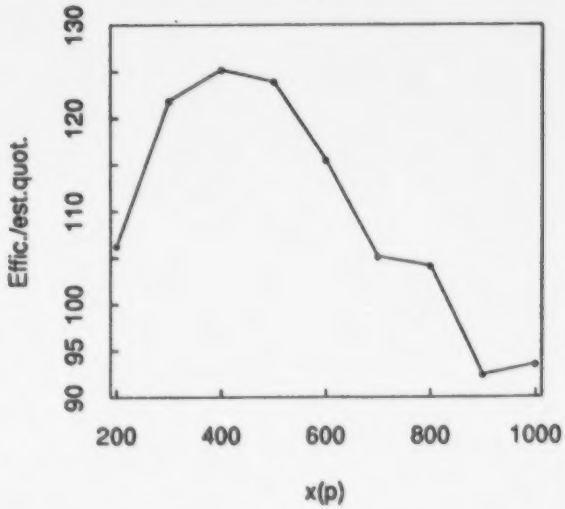


fig. 5: graphique de l'efficacité de $\widehat{\overline{Y}}_M$ par rapport à $\widehat{\overline{Y}}_q$.



On peut remarquer sur la figure 5 que l'efficacité de l'estimateur $\widehat{\overline{Y}}_M$ décroît assez rapidement lorsque la valeur de $x(p)$ augmente. Cette observation est prévue en théorie et va nous conduire à étudier un estimateur qui peut mieux se comporter dans de telles situations. L'idée qui a été suggérée par Mallows(1983) consiste à borner en plus de l'influence du résidu, l'influence de la variable explicative.

5. ESTMATEUR DE MALLOWS.

L'estimateur de Mallows du paramètre $\overline{Y}$ est donné par

$$\widehat{\overline{Y}}_{MA} = \widehat{\beta}_{MA}\,\overline{X}$$

où $\widehat{\beta}_{MA}$ est implicitement défini comme solution de l'équation

$$\sum_{i\in\theta} \psi_b\left(\frac{\sqrt{x_i}}{\mu}\right)\psi\left(\frac{y_i - \xi x_i}{\sqrt{x_i}}\right) = 0$$

$$\text{où} \quad \begin{cases} \mu = \textit{médiane de } \sqrt{X} \\ \psi_b = \textit{fonction de Huber} \\ \psi = \textit{fonction de robustesse quelconque.} \end{cases}$$

Il est à noter que μ est une constante de normalisation et la médiane a été choisie pour préserver la robustesse de la fonction ψ_b.

• CHOIX DE LA CONSTANTE b.

La constante b aura comme effet de limiter l'influence d'une proportion p de valeurs de X. En considérant X comme une variable aléatoire, cette proportion est donnée par la relation

$$p = P\left(\frac{\sqrt{X}}{\mu} \geq b\right)$$

Le tableau 2 ci dessous nous donne des valeurs de b pour quelques lois de probabilité et certaines valeurs de p.

	Valeur de p				
Loi de X	*5%*	*10%*	*15%*	*20%*	*25%*
$Log\,N(\mu,1)$	*2.2760*	*1.8979*	*1.6790*	*1.5232*	*1.4011*
$\lvert N(0,\sigma^2)\rvert$	*1.7046*	*1.5616*	*1.4609*	*1.3784*	*1.3059*
$exp(\lambda)$	*2.0789*	*1.8226*	*1.6544*	*1.5238*	*1.4142*
Table 2 : Valeurs de b comme fonction de p et de $\mathcal{L}(X)$					

Introduisons le paramètre β_{MA} défini par l'équation

$$\sum_{i=1}^{N} \psi_b\left(\frac{\sqrt{x_i}}{\mu}\right)\psi\left(\frac{y_i-\beta x_i}{\sqrt{x_i}}\right)=0 \qquad (6)$$

Il est possible d'établir pour une taille d'échantillon suffisamment élévée qu'on a:

$$\widehat{\beta}_{MA} \simeq \beta_{MA} + \frac{(1/n)\sum_{i\in s}\psi_b\left(\frac{\sqrt{x_i}}{\mu}\right)\psi\left(\frac{y_i-\beta_{MA}x_i}{\sqrt{x_i}}\right)}{(1/N)\sum_{i=1}^{N}\sqrt{x_i}\,\psi_b\left(\frac{\sqrt{x_i}}{\mu}\right)\psi'\left(\frac{y_i-\beta_{MA}x_i}{\sqrt{x_i}}\right)}$$

Cette équation nous permet d'établir que $\widehat{\bar{Y}}_{MA}$ est un estimateur biaisé et que son biais asymptotique est donné par:

$$B(\widehat{\bar{Y}}_{MA}) = \beta_{MA}\,\bar{X} - \bar{Y}$$

L'erreur quadratique moyenne de $\widehat{\bar{Y}}_{MA}$ est donnée par:

$$EQM(\widehat{\bar{Y}}_{MA}) = \left[B(\widehat{\bar{Y}}_{MA})\right]^2 + \left(\frac{\bar{X}}{\bar{\delta}}\right)^2 \frac{1-f}{n(N-1)}\sum_{i=1}^{N}\left[\psi_b\left(\frac{\sqrt{x_i}}{\mu}\right)\psi\left(\frac{y_i-\beta_{MA}x_i}{\sqrt{x_i}}\right)\right]^2$$

$$\text{où} \quad \begin{cases} \bar{\delta} = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N}\delta_i \\ \delta_i = \sqrt{x_i}\,\psi_b\left(\dfrac{\sqrt{x_i}}{\mu}\right)\psi'\left(\dfrac{y_i-\beta_{MA}x_i}{\sqrt{x_i}}\right) \end{cases}$$

Un estimateur de l'erreur quadratique moyenne de $\widehat{\bar{Y}}_{MA}$ est donné par:

$$eqm(\widehat{\bar{Y}}_{MA}) = \left[b(\widehat{\bar{Y}}_{MA})\right]^2 + \left(\frac{\bar{X}}{\bar{\delta}_s}\right)^2 \frac{1-f}{n(n-1)}\sum_{i\in s}\left[\psi_b\left(\frac{\sqrt{x_i}}{\mu}\right)\psi\left(\frac{y_i-\widehat{\beta}_{MA}x_i}{\sqrt{x_i}}\right)\right]^2$$

$$\text{où} \quad \begin{cases} \bar{\delta}_s = (1/n)\displaystyle\sum_{i\in s}\widehat{\delta}_i \\ \widehat{\delta}_i = \sqrt{x_i}\,\psi_b\left(\dfrac{\sqrt{x_i}}{\mu}\right)\psi'\left(\dfrac{y_i-\widehat{\beta}_{MA}x_i}{\sqrt{x_i}}\right) \end{cases}$$

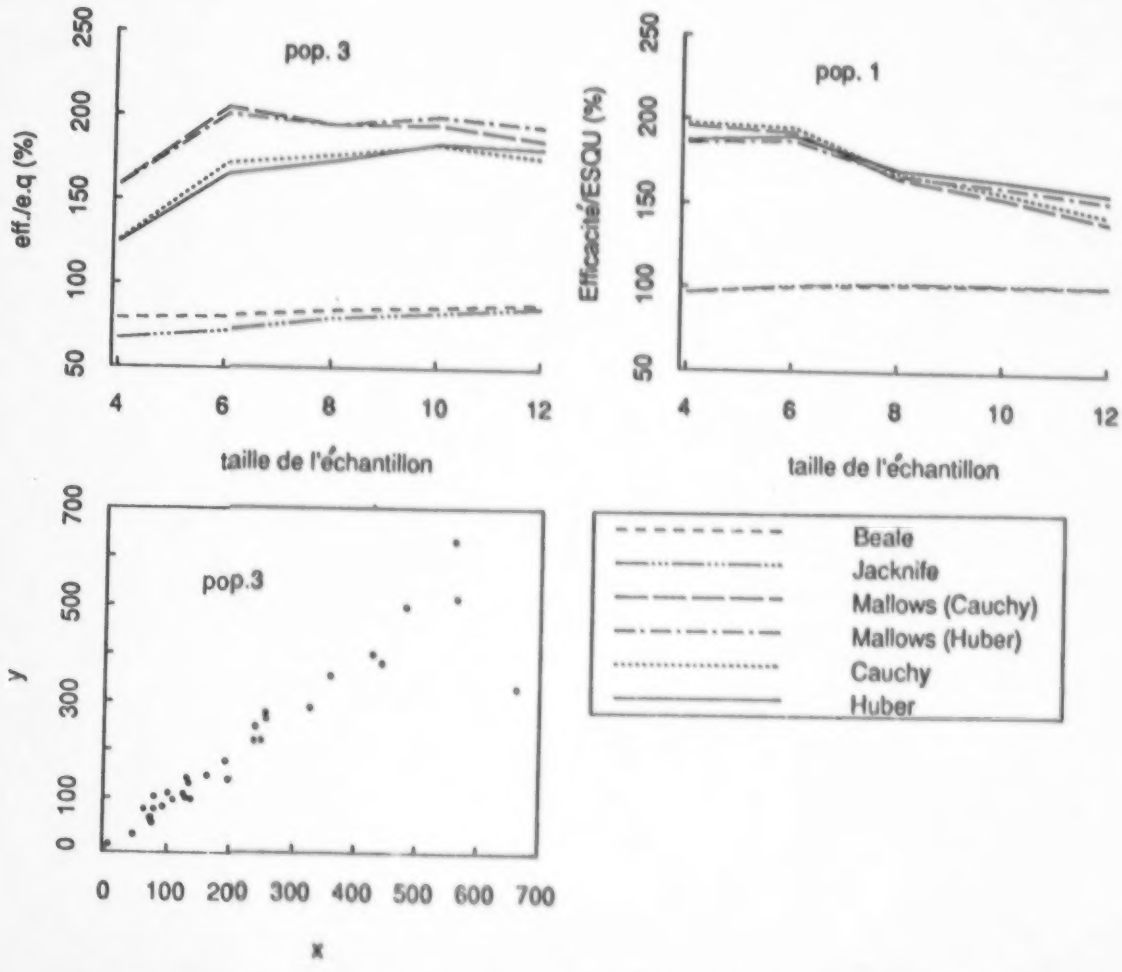
• Comme pour le M-estimateur, une taille d'échantillon n assez faible ne permettra pas d'avoir une estimation satisfaisante du biais; alors l'estimatiom de la variance constituera une meilleure estimation de l'erreur quadratique moyenne.
• Le calcul de l'estimateur $\widehat{\beta}_{MA}$ se fait aussi par la méthode itérative de Beaton & Tukey(1974).

6. ETUDE EXPÉRIMENTALE.

On considère la population de 34 villages de l'exemple 1. Dans un premier temps, Y représente la surface de blé en 1936 et X la surface totale cultivée en 1931 (Population 1). Dans un deuxième temps nous allons considérer comme variable d'intérêt Y la surface de blé en 1937 et comme variable auxilliaire X la surface de blé en 1936 (Population 3). Les deux graphiques d'observations sont représentés sur la figure 6. Dans chaque cas, 2000 échantillons aléatoires simples ont été tirés pour les tailles d'échantillons de 4, 6, 8, 10 et 12. Pour la fonction de cauchy, la constante c vaut $2.385\widehat{\sigma}$ et pour la fonction de Huber, $k = 1.345\widehat{\sigma}$. Les graphiques représentant l'efficacité des estimateurs étudiés par rapport à l'estimateur quotient sont également présentés sur la figure 6.

On peut y remarquer un gain important au niveau de l'efficacité résultant de l'utilisation des estimateurs robustes. La population 3, qui comporte une valeur aberrante ayant une grande valeur de X est un cas type où on peut s'attendre à ce que l'estimateur de Mallows soit le plus performant.

fig. 6: résultats de l'étude empirique



Bibliographie.

[1] Beaton, A. E. & Tukey, J. W.,(1974). " The Fitting of Power Series, Meaning Polynomials, Illustrated on Band-spectroscopic Data."
Technometrics, **16**, pp. 147-185.

[2] Brewer, K. W. R.,(1963b). " Ratio Estimation in Finite Populations: Some Results Deducible from the Assumption of an Underlying Stochastic Process."
Australian Journal of Statistics, **5**, pp. 93-105.

[3] Cochran, W. G.,(1977). *Sampling Techniques* (3^{th} *ed.*),
New York: John Wiley & Sons.

[4] Mallows, C. L.,(1983). " Discussion on Huber: Minimax Aspects of Bounded-Influence Regression."
Journal of the American Statistical Association, **78**, p. 77.

[5] Royall, R. M.,(1970a). " On Finite Population Sampling Theory under Certain Linear Regression Models."
Biometrika, **57**, pp. 377-387.

[6] Royall, R. M.,(1970b). " Finite Population Sampling - on Labels in Estimation."
Annals of Mathematical Statistics, **41**, pp. 1774-1779.

[7] Sukhatme, P. V. & Sukhatme, B. V.,(1970). *Sampling Theory of Surveys with Applications*.
Food and Agriculture Organization: Rome, second edition, p. 185.

[8] Wu, C. F.,(1982)."Estimation of Variance of Ratio Estimator."
Biometrika, **69**, pp. 183-189.

PC CARP: Un logiciel pour le calcul de la variance dans les enquêtes complexes[1]

Michel A. Hidiroglou[2]

Les grandes enquêtes sont, en général, stratifiées et échantillonnées à plusieurs degrés. Les données de telles enquêtes sont analysées en se servant d'une batterie d'analyses statistiques qui comportent le calcul de moyennes, de totaux, de rapports, ainsi que de la régression simple et logistique. De telles analyses, bien souvent, ne tiennent pas compte du plan de sondage. Cet article met en lumière les principaux aspects de PC CARP, un logiciel pour l'ordinateur personnel qui tient compte du plan de sondage.

1. INTRODUCTION

Les grandes enquêtes sont, en général, stratifiées et échantillonnées à plusieurs degrés. De tels plans de sondages sont favorisés par les grands centres de sondage parce qu'ils offrent plusieurs avantages au niveau opérationnel et au niveau des coûts. Cependant, l'analyse des données et en particulier le calcul de variances de tels sondages est considérablement plus compliqué que pour celui associé aux échantillons aléatoires simples. Les logiciels tels que SAS, BMDP et SPSS ne tiennent jusqu'à présent pas comptent du plan de sondage en calculant les variances de sondages complexes.

Au début des années 70, l'université d'Iowa State, se fondant sur les travaux d'Hidiroglou (1974) et de Fuller (1975), élaborait un programme ayant pour but d'estimer des coefficients de régression et la matrice des covariances de ces coefficients pour des données d'enquête. Ce programme, connu sous le nom de SUPER CARP, devait plus tard être modifié de manière à inclure l'estimation d'agrégats et de rapports, les statistiques de sous-populations, les tableaux de contingences à doubles entrées et les échantillons à deux degrés. La dernière revision de SUPER CARP remonte à 1980. Ce programme a ouvert la voie à la création d'un logiciel pour ordinateur personnel. Ce dernier, nommé PC CARP pour des raisons de continuité, a

---

[1]) Présenté au colloque sur les méthodes et domaines d'application de la statistique du 57è congrès de l'Association Canadienne Française pour l'Avancement des Sciences, UQUAM, mai 1989.

[2]) Division des méthodes d'enquêtes-entreprises, Statistique Canada, Ottawa, Ontario, K1A 0T6

été développé à l'université d'Iowa State en collaboration avec le International Statistical Programs Center du U.S. Census Bureau. Il faisait partie d'une gamme de cinq logiciels pour ordinateur personnel conçus afin que les pays en voie de développement puissent traiter et analyser leurs données sur place.

Ce document donne plusieurs points saillants de ce logiciel. Pour de plus amples renseignements, le lecteur peut lire l'article par Snell et coll. (1988).

2. DESCRIPTION DU PROGRAMME

PC CARP peut accommoder avec autant de facilité de petits ou de grands ensembles de données. Il n'y a aucune restriction sur le nombre de strates ou de grappes qui peuvent être inclues dans les analyses. Cependant, il ne traite que 50 variables à la fois pour chaque analyse.

Afin que le programme puisse reconnaître la stratification ainsi que le regroupement des données à l'intérieur des grappes, il est important qu'elles soient ordonnées par strate et par grappe (selon le nombre de degrés représentés). Pour incorporer la correction pour la population finie, si nécessaire, des facteurs les représentant sont aussi lus par le programme. Les données sont identifiées comme suit: la strate, la grappe du premier degré (unité primaire d'échantillonnage) suivies de l'identification des grappes des degrés secondaires, les données, et le poids des observations. Le poids correspond à l'inverse de la sélection globale, en tenant compte de la stratification ainsi que du nombre de degrés associés au plan. Donc pour des fins strictement de définition, chaque observation sera identifiée par la strate "h", la grappe "i", les données représentées par le vecteur x_{hik}, et le poids par w_{hik} où k est la $k^{ième}$ unité observée.

On peut diviser le genre d'analyse en trois groupes: simple, intermédiaire et complexe. Les analyses simples comprennent des totaux, rapports, différences entre deux rapports, des totaux par strate, des moyennes par strate et des proportions par strate. Les analyses intermédiaires ajoutent la dimension "sous-population" aux analyses simples. Les analyses complexes incluent les tableaux de contingence, la régression simple et logistique ainsi que l'analyse univariée. Pour toutes ces analyses, le programme calcule des variances. Pour certaines, il obtient des coefficients de variation, des matrices de covariances et l'effet du plan de sondage. Le tableau 1

nous donne un sommaire de ces options. L'effet du plan de sondage, inventé par Kish (1965), et connu sous la notation DEFF, est tout simplement le rapport entre la variance calculée selon le plan de sondage et la variance que l'on aurait obtenue si le plan de sondage avait été aléatoire simple. Le DEFF nous donne une très bonne indication de combien l'on doit modifier les résultats d'analyses qui auraient ignoré le plan de sondage.

Tableau 1
Fonctions d'analyse disponibles dans PC CARP

Analyse	Coeff-cient de variation	Matrice des co-variances	Effet du plan	Remarques
Analyses de populations				
Estimation d'agrégats	X	X	X	50 variables au maximum
Estimation de rapports	X	X	X	50 variables au maximum sans les covariances, 15 avec covariances
Différence de rapports			X	15 variables au maximum
Analyses de strates				
Agrégats	X		X	50 variables au maximum
Moyennes	X		X	50 variables au maximum
Proportions	X		X	50 variables au maximum
Analyses de sous-populations				
Agrégats	X		X	Classement recoupé Variables multiples
Moyennes	X		X	Classement recoupé Variables multiples
Proportions	X		X	Classement recoupé Variables multiples
Rapports	X		X	Classement recoupé Variables multiples
Autres analyses				
Tableau à double entrée		X		50 cases au maximum, test de proportionnalité
Régression		X		50 variables au maximum Tests à plusieurs degrés de liberté, valeurs prédites résidus
Analyse unidimensionnelle			X	Variables multiples, fonctions de distribution cumulatives empiriques, quantiles

On remarquera que dans le cas des analyses de sous-populations, le programme offre la possibilité à l'utilisateur de croiser des variables de classification à partir de classifications existantes. Ceci permet la construction de tableaux de contingence multi-dimensionnelles. L'analyse de la régression par moindres carrés pondérés est basée sur les calculs présentés dans Hidiroglou (1974) et Fuller (1975). Le test chi-carré pour les l'hypothèses d'indépendance dans les tableaux de contingence est basé sur les calculs de Rao et Scott (1984).

3. REMARQUES OPÉRATIONNELLES

PC CARP fonctionne sur un ordinateur du type IBM PC ou IBM PC/XT, et exige une mémoire minimum de 410 kilo-octets. Il faut aussi que l'ordinateur soit muni d'un coprocesseur "mathématique", afin de permettre le calcul de certaines fonctions.

La communication entre l'utilisateur et l'opérateur est uniquement faite à l'aide d'écrans interactifs qui posent une série de questions spécifiques. Pour certaines questions, une réponse par défaut peut être modifiée par l'utilisateur. Les mauvaises réponses sont rejetées par le programme. Le système de traitement d'erreurs a été conçu de manière à prévenir des arrêts de programme causés par des erreurs de l'utilisateur. Le programme renferme des fonctions de vérification pour l'absence de réponse, l'inexactitude des noms de fichiers et l'inexactitude des définitions des variables d'analyse. Si une erreur est décélée, PC CARP permet à l'utilisateur de réintroduire les données ou de sortir du programme.

PC CARP a une série de menus où les questions sont posées. Il y a deux groupes de menus principaux. Le premier est la SPÉCIFICATION DU PROBLÈME. On s'en sert afin d'identifier les variables d'analyses, de nommer ces variables, et de rentrer dans le programme les données ainsi que leurs identificateurs. Le deuxième menu principal s'appelle LA SPÉCIFICATION DE L'ANALYSE. Il sert à spécifier l'analyse désirée, identifier les variables participant à l'analyse, et aussi à fournir à l'ordinateur les commandes nécessaires pour effectuer l'analyse.

La SPÉCIFICATION DU PROBLÈME est divisée en cinq menus. Ceux-ci servent à identifier l'analyse, le nom des variables, le type de sondage, le format des données, ainsi que la destination pour la sortie de l'analyse. Pour certains menus, il y a plus d'un écran. Afin de corriger des menus précédents, une commande permet

le retraçage. Elle permet à l'utilisateur de revenir en arrière sans affecter l'exécution du programme. Si la question n'est pas claire à l'utilisateur, il peut demander des précisions en évoquant la commande "HELP". Cette commande va lui permettre de mieux comprendre la question en lui donnant des réponses spécifiques ou générales. Les résultats peuvent sortir soit sur l'écran, soit sur papier, ou être mis en mémoire dans l'ordinateur. La sortie des résultats est divisée en deux parties. La première partie est obtenue après la phase de la SPÉCIFICATION DU PROBLÈME et consiste en un sommaire de la configuration des données spécifiées par l'utilisateur. La deuxième partie consiste en statistiques calculées pour chaque analyse spécifiée. Plusieurs analyses peuvent être effectuées une fois que la phase de la spécification du problème a été faite. La sortie de ces analyses varie selon l'analyse spécifiée.

PC CARP n'accepte que des ensembles de données sans valeurs manquantes. Un programme auxiliaire, appelé PRE CARP, peut servir à remplacer les données manquantes.

4. CALCULS

Le programme est très efficace pour le calcul des statistiques. Il consiste en un algorithme monopasse servant à calculer les statistiques élémentaires ainsi que leurs variances. L'algorithme choisi est celui de Herraman (1968). Étant donné les valeurs d'échantillon x_i ainsi que leurs poids correspondants w_i, les moyennes pondérées $\bar{x}_k$, et les sommes des carrés corrigées pondérées S_K sont calculées à l'aide des formules suivantes.

$$\bar{x}_k = \bar{x}_{k-1} + a_k d_k \text{ et } S_k = S_{k-1} + D_k - D_k a_k,$$

où

$$d_k = x_k - \bar{x}_{k-1},\ a_k = w_k \left(\Sigma_{i=1}^{k} w_i\right)^{-1},\ \text{et } D_k = d_k^2 w_k.$$

Pour des analyses plus complexes, telles que la régression, l'analyse univariée ou le calcul de plan à plusieurs degrés, il est nécessaire de faire deux passes sur les données.

Le calcul de variance pour les fonctions non-linéaires des moments pondérées de l'échantillon se fait en se servant de la linéarisation de Taylor. Cette méthode est bien supportée du point de vue théorique par les articles de Woodruff (1971),

Kish et Frankel (1974), Fuller (1975), ainsi que par les livres de Cochran (1977) et de Wolter (1985).

Cette méthode exprime une statistique t comme une fonction $g(y)$ de $y = (Y_1, \ldots, Y_p)$ ou Y_j est le total de la variable "j" pour la population. Un estimateur consistant de $g(y)$ est $g(\hat{y})$ ou $\hat{y} = (\hat{Y}_1, \ldots, \hat{Y}_p)$ est l'estimateur de y. Étant donné que $\hat{Y}_j = \Sigma_{h=1}^{L} \Sigma_{i=1}^{n_h} w_{hi}\, y_{hij}$, on trouve que

$$\mathrm{Var}[g(\hat{y})] = \Sigma_{h=1}^{L} V[\Sigma_{i=1}^{n_h} w_{hi} \hat{z}_{hi}],$$

où

$$\hat{z}_{hi} = \Sigma_{j=1}^{p} y_{hij}\, g^{(j)}(\hat{y}),\quad g^{(j)}(\hat{y}) = \left.\frac{\partial g(y)}{\partial Y_j}\right|_{Y_j = \hat{Y}_j},$$

h est la strate, et i est l'unité primaire d'échantillonnage.

Un estimateur consistant de $\mathrm{Var}[g(\hat{y})]$ est $\Sigma_{h=1}^{L} v(\hat{z}_h)$ où $\hat{z}_h = \Sigma_{i=1}^{n_h} w_{hi} \hat{z}_{hi}$.

Si nous prenons le cas de l'estimation d'un rapport pour un sondage stratifié aléatoire simple, l'estimateur est

$$r = \sum_{h=1}^{L} \sum_{i=1}^{n_h} w_{hi}\, y_{hi} \Big/ \sum_{h=1}^{L} \sum_{i=1}^{n_h} w_{hi}\, x_{hi}$$

et son estimateur de variance sous forme linéaire est

$$v(r) = \sum_{h=1}^{L} (1-f_h)\, n_h\, (n_h - 1)^{-1} \sum_{i=1}^{n_h} (z_{hi} - \bar{z}_h)^2 \Big/ \hat{X}^2$$

où $z_{hi} = w_{hi}\,(y_{hi} - r\, x_{hi})$ et $\bar{z}_h = \Sigma_i\, z_{hi}/n_h$.

Il est à noter que PC CARP peut aussi calculer des variances reflétant la stratification à postériori en se servant de la linéarisation de Taylor. Nous allons démontrer ces calculs en se servant de l'estimateur à postériori de l'enquête sur la population active au Canada. L'estimateur est de la forme

$$\hat{Y}_{post} = \Sigma_a\, [({}_aN/{}_a\hat{N})\, {}_a\hat{Y}]$$

où ${}_a\hat{Y} = \Sigma_h\Sigma_i\Sigma_k w_{hik}\ {}_a y_{hik}$, ${}_a\hat{N} = \Sigma_h\Sigma_i\Sigma_k w_{hik}\ {}_a 1_{hik}$,
et ${}_a 1_{hik}$ est une variable indicatrice égale à un si la hik à unité appartient à la post-strate et zéro autrement; ${}_a y_{hik} = y_{hik}\ {}_a 1_{hik}$ et ${}_a N$ $(a=1,\ldots,A)$ sont des comptes connus de la population classés par age et sexe.

Si l'on suppose que les grappes "i" ont été tirées avec remise dans chaque strate, l'estimateur de variance pour $\hat{Y}$ est

$$v({}_a\hat{Y}) = \Sigma_h\ n_h(n_h-1)^{-1}\ \Sigma_i\ (z_{hi} - \bar{z}_h)^2$$

où $z_{hi} = \Sigma_k\ w_{hik}\ [y_{hik} - \Sigma_a\ ({}_a\hat{Y}/{}_a\hat{N})\,{}_a 1_{hik}]$

et $\bar{z}_h = n_h^{-1}\ \Sigma_i\ z_{hi}$.

Le programme est conçu pour ne pas calculer des variances dans les strates où il n'y a qu'une seule observation. La variance est mise à zéro et un message d'avertissement est émis. Pour ce genre de strates, l'utilisateur peut choisir de les regrouper automatiquement avec des strates adjacentes. Si l'on dénote les taux de sondage des strates par f, les nouveaux taux sont définis par

$$f_h^* = (n_h\ f_h^{-1} + n_{h+1}\ f_{h+1}^{-1})^{-1}(n_h+n_{h+1})$$

où la strate h, avec $n_h=1$, a été combinée avec la strate h+1. L'ordre de classement des strates influera sur le mode de groupement des séries de données et ainsi sur les taux de sondage des strates. Le programme exige une lecture additionnelle lorsque l'utilisateur a recourt à l'option "regroupement de strates" ou à l'option "à deux degrés".

La précision des calculs effectués a été évalué en se servant des données de Longley (1967). On a constaté que PC CARP avait une précision numérique équivalente à celle des principaux progiciels comme SAS.

5. CONCLUSION

Il y a présentement au delà de 100 exemplaires de PC CARP répartis dans plus de 20 pays. Ce logiciel a facilité aux rechercheurs les calculs reflétant le plan de sondage pour une gamme assez importante d'estimateurs. Le logiciel sera continuellement amélioré en lui ajoutant d'autres procédures. Cependant pour le moment, il contient assez d'options pour satisfaire les besoins de la majorité des utilisateurs. Sa grande facilité d'utilisation est aussi un grand avantage.

On peut se procurer une copie du logiciel pour une somme nominale en écrivant à

Professor Wayne A. Fuller
Department of Statistics
Room 204, Snedecor Hall
Iowa State University
Ames, Iowa, 50011

Bibliographie

. Cochran, W.G. (1977). *Sampling Techniques.* New York: John Wiley.

. Fuller, W.A. (1975). Regression analysis for sample survey. *Sankhyā* **C 37**, 117-132.

. Herraman, C. (1968). Sums of squares and products matrix. *Applied Statistics*, **17**, 289-292.

. Hidiroglou, M.A. (1974). Estimation of regression parameters for finite populations. Thèse de doctorat non publiée. Iowa State University, Ames, Iowa.

. Kish, L. (1965). *Survey Sampling.* New York: John Wiley.

. Kish, L. et Frankel, M.R. (1974). Inference from complex samples. *Journal of the Royal Statistical Society*, B, 36, 1-37.

. Longley, J.W. (1967). An appraisal of least squares programs for the electronic computer from the point of view of the user. Journal of the American Statistical Association, **62**, 819-841.

. Rao, J.N.K, et Scott, A.J. (1984). On chi-square tests for multiway contingency tables with cell proportions estimated from survey data. *Annals of Statistics* **12**, 46-60.

. Snell, D., Kennedy, W.J., Sullivan, G., Park, H.J. et Fuller, W.A., (1988). Logiciel d'ordinateur personnel pour l'estimation de la variance dans les enquêtes complexes. *Techniques d'enquête*, Vol. **14**, no. 1, 59-69.

. Wolter, K.M. (1985). *Introduction to Variance Estimation.* New York: Springer-Verlag.

. Woodruff, R.S. (1971). A Simple Method for Approximating the Variance of a Complicated Estimate. *Journal of the American Statistical Association*, **66**, 411-414.

ÉVALUATION DE LA MÉTHODE DE SÉLECTION DE L'ÉCHANTILLON AU RECENSEMENT DE LA POPULATION CANADIENNE

Marc Joncas
Statistique Canada
Juin 1989

MOTS-CLÉS

Représentativité, Chi-deux, Student, Hypergéométrique, biais.

RÉSUMÉ

Un biais dans la méthode de sélection de l'échantillon est une possibilité dont on doit toujours se méfier lors du recensement de la population canadiennne. De fait, un tel biais s'il existe, peut avoir de sérieux effets sur les estimations du recensement (distributions, projections intercensitaires,...). Il est, par conséquent, important de se doter d'outils nous permettant de détecter la présence d'un biais. Plusieurs méthodes ont été conçues pour parvenir à cet objectif. Cet article décrit quelques unes des méthodes utilisées (ou envisagées) par Statistique Canada. Des discussions sont données pour identifier les points forts et les points faibles de chacune des méthodes présentées.

0.1 INTRODUCTION

De nos jours, l'échantillonnage constitue une technique de plus en plus utilisée dans maints domaines. Il est ainsi possible d'obtenir une quantité substantielle d'informations sur des sujets très variés à un coût moindre que ceux engendrés avec la mise en place d'un recensement par exemple. Mais souvent, puisque l'idée est lancée, un doute persiste quant à la pertinence des résultats obtenus à partir de cette fraction seulement de la population. Dans le cadre de cette présentation, nous ferons un tour d'horizon des méthodes les plus couramment utilisées à Statistique Canada en vue d'évaluer le mode de sélection de l'échantillon lors des recensements de la population canadienne. De plus, nous introduirons deux méthodes qui seront proposées dans le cadre des études s'inscrivant au recensement de 1991.

0.2 CONTEXTE DE TRAVAIL

Lors d'un recensement de la population, la plupart des strates géographiques font l'objet d'un échantillonnage systématique des ménages qu'elles abritent (*). Ainsi, deux types de questionnaires sont distribués; 4 cinquième des ménages reçoivent un questionnaire court comportant des questions de base (âge, sexe,...) et l'autre cinquième reçoit un questionnaire long qui, en plus des mêmes questions de base (on parle alors d'un recensement complet de ces questions), comporte des questions supplémentaires (revenu, profession,...) (on parle alors d'un échantillonnage des questions supplémentaires). La question légitime que l'on se pose est la suivante : pour les questions faisant l'objet d'un dénombrement

complet, ou de façon équivalente, pour les questions communes aux deux types de questionnaires, l'échantillon obtenu donne-t-il une image acceptable de la population? Dans ce qui suit, plusieurs éléments de réponses à cette question sont présentés.

(*) Pour des raisons d'ordre opérationnel, il est préférable de recenser tous les ménages de certaines strates géographiques. De plus, on remarque que le terme 'strate géographique' est utilisé à la place du terme 'secteur de dénombrement' connu seulement des utilisateurs des données du recensement.

1.1 PREMIÈRE MÉTHODE

Les variables communes aux deux types de questionnaires sont regroupées selon qu'elles soient des variables ménages c.-à-d. issues des questions se rapportant aux ménages seulement, ou par opposition, des variables population c.-à-d. issues des questions se rapportant aux individus comme tels. Cette première méthode a été utilisée en 1976 [1] afin d'évaluer l'hypothèse nulle d'une bonne représentativité de l'échantillon des ménages du recensement. On a alors considéré la statistique suivante pour chacune des variables ménages à l'étude.

$$t(p_i) = (p_i - P_i)/\sqrt{s_i^2(1-f_i)/n_i},$$

où f_i correspond à la fraction de sondage dans la strate i à l'étude,

p_i donne la proportion (non pondérée) de ménages de l'échantillon ayant la caractéristique d'intérêt,

P_i donne la proportion de ménages dans la population ayant la caractéristique d'intérêt,

$s_i^2 = n_i p_i(1-p_i)/(n_i-1)$ et

n_i = le nombre de ménages échantillonnés dans la strate i.

Alors si on suppose que les p_i sont distribués selon une loi normale, et que les ménages de l'échantillon ont été sélectionnés selon un échantillonnage aléatoire simple, nous avons que la statistique t(p) suit une distribution de Student avec n -1 degrés de liberté.

1.2 MÉCANIQUE DU TEST

Le déroulement du test se résume à ces quatre étapes.

1- Calculer $d_i = Pr(|t(p_i)| \geq |t(p_i^*)|)$ la probabilité d'obtenir un résultat supérieur ou égal au résultat p_i^* observé à partir de l'échantillon selon une distribution de Student.

2- Obtenir le nombre f^* de strates pour lesquelles d_i est inférieur à une valeur d'un point critique Pc déterminé (par ex. Pc=5%).

3- Calculer la probabilité h, d'obtenir un résultat égal ou supérieur à f^* c.-à-d.

$$h=Pr(\text{le nombre de résultats significatifs} \geq f^*)$$
$$=\sum_{g=f^*}^{N}\binom{N}{g}(Pc)^g(1-Pc)^{N-g}$$

où N donne le nombre de strates à l'étude.

4- Rejeter l'hypothèse d'une bonne représentativité de l'échantillon si la valeur de h est jugée trop petite (inférieur à 5% par exemple).

1.3 DISCUSSION

Ici on peut toujours s'interroger sur les implications d'un rejet de l'hypothèse nulle d'une bonne représentativité de l'échantillon ;

a)on peut rejeter à tort l'hypothèse nulle (généralement on accepte de rejeter en moyenne 5% des fois un échantillon pour lequel l'hypothèse nulle est vérifiée);

b)il se peut que les hypothèses nous permettant de conclure que la statistique t(p) suive une distribution de Student sous l'hypothèse nulle soient violées.

c)on peut également rejeter l'hypothèse nulle à cause d'un biais dans le mode de sélection de l'échantillon.

d)il se peut que l'on rejette l'hypothèse nulle à cause d 'une combinaison des raisons données aux points b) et c).

1.4 AVANTAGES ET DÉSAVANTAGES

Cette méthode a le double avantage d'être facile à comprendre et à utiliser. Cependant elle comporte certaines lacunes. En premier lieu, on remarque qu'elle ne s'applique qu'à une catégorie de la variable à l'étude. Par exemple, si on s'intéresse à la représentativité de la langue maternelle du chef de famille, et que cette dernière compte 10 catégories différentes (français, anglais, italien,...) il nous faudra calculer cette statistique (t(p)) pour chacune des catégories. Cette méthode ne nous rassure donc pas sur la qualité de l'échantillon pour l'ensemble des catégories. De plus, cette méthode n'est valide en principe que pour des échantillons aléatoires simples. Or en pratique, on est en présence d'un échantillon systématique de ménages. En troisième lieu, cette méthode est très sensible au fait que l'on puisse supposer (à l'aide du théorème de la limite centrale par exemple) que les p soient répartis selon une loi normale.

Pour illustrer ce dernier point, étudions l'exemple suivant: supposons que M la taille de la strate soit de 100 ménages, que X le nombre de ménages ayant la caractéristique d'intérêt soit de 10 et que la taille de l'échantillon soit de 20. De plus, supposons que les variables soient de type dichotomique. Alors la probabilité $Pr(|p-P|\geq c)$ (*) peut être calculée de façon exacte sous l'hypothèse d'un échantillon aléatoire simple sans remise en utilisant la distribution hypergéométrique [2]. Ces probabilités apparaissent dans le tableau I

de même que les probabilités calculées à l'aide d'une loi de Student.

(*) On peut montrer qu'il existe une valeur unique de c telle que la probabilité $Pr(|p-P| \geq c)$ soit équivalente à la probabilité $Pr(|t(p_i)| \geq |t(p_i^*|)$. La simplicité d'écriture de la première expression a guidé ce changement de notation.

TABLEAU I. PROBABILITÉS D'OBSERVER UN ÉCART (p-P) SUPÉRIEUR OU ÉGAL À UNE VALEUR c DONNÉE.

	Pr(|p-P|≥c)	
c	Hypergéométrique	Student
0	1.00000	1.00000
.05	0.47714	0.68279
.10	0.17922	0.18922
.15	0.02153	0.08048
.20	0.00354	0.03567
.25	0.00037	0.01514

En supposant une distribution de Student, Pr(|p-P|>=0.20) = 0.03567. Donc, tout échantillon pour lequel |p-P|>=0.20 est considéré comme étant significatif à un niveau alpha = 5%. Par exemple, si le nombre de strates à l'étude est de 1000, alors on devrait s'attendre à retrouver en moyenne 36 (0.03567*1000) strates significatives. Or puisque la probabilité réelle de trouver un résultat tel que |p-P|>=0.20 est de 0.00354 (selon la loi hypergéométrique) le nombre de strates significatives que l'on devrait retrouver serait plutôt de 4 (i.e. 0.00354*1000). En conclusion, à moins que nous nous trouvions en présence d'un cas d'extrême biais, le fait d'avoir recours à l'approximation de Student nous amène à accepter très souvent l'hypothèse nulle qui stipule qu'il n'y a pas de biais de représentativité de l'échantillon.

2.1 DEUXIÈME MÉTHODE

Lorsque la taille de l'échantillon est faible ou lorsque l'approximation de la distribution ne peut s'apparenter à une distribution de Student, on utilise la loi hypergéométrique (c.-à-d. la loi exacte). Son utilisation suppose cependant que nous soyons en présence de données discrètes et d'un échantillon aléatoire simple. Encore une fois, cette méthode ne s'applique que pour l'étude des variables ménages.

2.2 MÉCANIQUE DU TEST

La façon de procéder avec cette deuxième méthode est exactement la même que celle décrite pour la première à l'exception du fait que le calcul des probabilités se fait à l'aide d'une loi hypergéométrique.

Ici cependant, étant donné la nature discrète des données, il semble inadéquat de poser Pc=5%. Si on revient à l'exemple de la section 1.4, on se rappelle que la probabilité d'observer un résultat significatif nous donnait une valeur de Pc = 0.02153. Cette différence peut avoir un impact important concernant la décision à rendre par rapport à la représentativité de l'échantillon. Par exemple, si nous avons 30000 strates à l'étude, et si Pc est supposé égal à 5%, on pourrait s'attendre à ce qu'en moyenne 1500 (c.-à-d. 30000*0.05) strates aient un résultat significatif. Or si f^*, le nombre de strates significatives, est effectivement de 1500 on a que

h(1500)=0.5000 si Pc est égal à 5.0%
h(1500)=0.0066 si Pc est égal à 4.7%

Donc, si Pc = 5%, on accepte l'échantillon comme étant représentatif de la population des ménages, alors que la décision est renversée si Pc est égal à 4.7%. En d'autres mots, un léger biais dans l'estimation de la valeur de Pc peut avoir un impact important sur la décision finale. En vue d'enrayer ce problème la solution suivante est proposée.

Si $d_i' = \max \Pr(|p_i - P_i| \geq c) < 5\%$ (on pose $d_i'=0$ si aucune valeur < 5%)

et si $G_i = \begin{cases} 0 \text{ si } d_i' \geq 5\% \\ 1 \text{ si } d_i' < 5\% \end{cases}$

$$\text{alors } E\left(\sum_i^N G_i/N\right) = 1/N \sum_i^N E(G_i) = 1/N \sum_i^N d_i' = \bar{d}'$$

$$V\left(\sum_i^N G_i/N\right) = 1/N^2 \sum_i^N V(G_i) = 1/N^2 \sum_i^N d_i'(1-d_i').$$

Si on pose maintenant $f = \sum_i^N G_i$, alors par le théorème de la limite centrale on a que

$$\frac{f/N - E(f/N)}{\sqrt{V(f/N)}} = \frac{f/N - \bar{d}'}{1/N\sqrt{\sum_i d_i'(1-d_i')}} \sim N(0,1)$$

et

$$h = \Pr\left(\frac{f - N\bar{d}'}{\sqrt{\sum_i d_i'(1-d_i')}} \geq \frac{f^* - N\bar{d}'}{\sqrt{\sum_i d_i'(1-d_i')}} \right) \sim N(0,1) .$$

La décision finale serait de rejeter l'hypothèse d'une bonne représentativité de l'échantillon vis-à-vis de cette variable si h<5%.

2.3 DISCUSSION

Les mêmes remarques faites lors de la présentation de la première méthode s'appliquent ici bien qu'on ait réduit de beaucoup la violation possible des hypothèses nous permettant d'utiliser la distribution hypergéométrique. Jusqu'à maintenant, le fait que l'échantillonnage en soit un systématique plutôt qu'aléatoire simple nous a limité à ne tester que les variables ménages (le plan de sondage des variables population avec utilisation de grappes de sélection s'éloignant beaucoup du modèle aléatoire simple désiré). De plus, les méthodes présentées étaient valides dans des conditions très précises (en termes de distribution des lois statistiques) et devaient être calculées autant de fois qu'il y avait de modalités pour la variable à l'étude. Nous présentons maintenant deux méthodes qui ont l'avantage d'adoucir ces restrictions. Il est à noter que ces méthodes n'ont pas été mises à l'essai au recensement de la population canadienne.

3.1 TROISIÈME MÉTHODE

La méthode proposée ici est égalememt connue sous le nom du test du Chi-deux généralisé de Wald [3].

$$X_W^2 = n_i(\hat{\underset{\sim}{p}}-\underset{\sim}{P}_0)'\underset{\sim}{V}^{-1}(\hat{\underset{\sim}{p}}-\underset{\sim}{P}_0)$$

où $\hat{\underset{\sim}{p}}$ et $\underset{\sim}{P}_0$ sont des vecteurs de dimension k-1 et $\underset{\sim}{V}$, une matrice carrée de variance-covariance de même dimension.

3.2 IDÉE INTUITIVE

Supposons que la population est divisée en k catégories, en accord avec les valeurs de la variable d'intérêt. Les proportions observées dans la population pour la strate à l'étude sont $P_1, P_2, \ldots P_k$ avec la contrainte usuelle $\sum P_j = 1$. Si on pose $\underset{\sim}{P} = (P_1, \ldots P_{k-1})'$, l'hypothèse nulle d'une bonne représentativité de l'échantillon devient

$$H_0 : \underset{\sim}{P}-\underset{\sim}{P}_0 = (P_{01}, \ldots P_{0k-1})'$$

Si $\hat{p}_1, \ldots \hat{p}_k$ sont des <u>estimations non biaisées</u> nous provenant de l'échantillon, alors on a que la statistique standard du Chi-deux devient

$$X_S^2 = n_i \sum(\hat{p}_j - P_{0j})^2/P_{0j}$$

On remarque ici que si on pose $\underset{\sim}{B}_0 = \text{Diag}(\underset{\sim}{P}_0) - \underset{\sim}{P}_0\underset{\sim}{P}_0'$ (c.-à-d. la matrice de variance-covariance de $\underset{\sim}{p}$ pour un échantillon indépendant lorsque H_0 est vraie) alors la statistique standard du Chi-deux peut s'écrire sous la forme

$$X_S^2 = n_i(\hat{\underset{\sim}{p}}-\underset{\sim}{P}_0)'\underset{\sim}{B}_0^{-1}(\hat{\underset{\sim}{p}}-\underset{\sim}{P}_0) \ .$$

Le problème avec cette statistique est que dans la plupart des enquêtes d'envergure, les plans d'échantillonnage consistent à stratifier les données de l'enquête, à sélectionner les données selon

plusieurs niveaux différents, et de ce fait, il n'est pas souvent très réaliste de considérer les observations comme provenant d'une distribution multinomiale. Cependant on pourrait montrer qu'à l'aide du théorème de la limite centrale, l'expression suivante

$$\sqrt{n_i}(\hat{\underset{\sim}{p}}-\underset{\sim}{P}) \dashrightarrow N(0,\underset{\sim}{V})$$

où $\underset{\sim}{V}$ est une matrice de variance-covariance définie semi-positive. Ainsi, si nous disposons d'une estimation de $\underset{\sim}{V}$ qui converge, disons $\hat{\underset{\sim}{V}}$, alors on peut tester l'hypothèse H_0 à l'aide de la statistique de Wald X_W^2 qui suit sous H_0 une distribution de Student avec k-1 degrés de liberté.

3.3 MÉCANIQUE DU TEST

La mécanique de ce test est tout à fait similaire aux mécaniques décrites pour les méthodes précédentes.

3.4 AVANTAGES ET DÉSAVANTAGES

Cette méthode est intéressante parce qu'elle nous permet de tester la représentativité de notre échantillon même si celui-ci nous provient d'une enquête au plan de sondage complexe (sans mentionner que l'on teste toutes les modalités de la variable en une seule fois). Le gros désavantage de cette approche cependant, est qu'il n'est pas toujours facile d'obtenir la matrice de variance-covariance $\underset{\sim}{V}$ (généralement très couteuse). De plus, on peut montrer que la statistique X_W^2 devient instable pour des valeurs de $\hat{p}_i$ qui seraient près de 1 ou 0, ou lorsque le nombre de degrés de liberté sur lequel est basée l'estimation de $\underset{\sim}{V}$ n'est pas relativement important comparativement au nombre de catégories de la variable d'intérêt. Ce qui nous amène à parler du dernier des tests présentés ici.

4.1 QUATRIÈME MÉTHODE

La dernière méthode discutée dans ce rapport réfère à la statistique modifiée du Chi-deux standard [4].

$$X_m^2 = X_S^2 * (k-1)/(\sum^{k} \hat{d}_j(1-P_j)$$

où $\hat{d}_j$: effet du plan de sondage par rapport à un échantillonnage aléatoire simple avec remise (é.a.s.a.r.)

4.2 IDÉE INTUITIVE

L'objectif de cette quatrième méthode consiste à adapter la statistique standard du Chi-deux X_S^2, de façon à ce que l'on puisse se servir des résultats ainsi obtenus aux conclusions à porter sur des données provenant d'enquêtes complexes. Alors on a que

$$X_S^2 = \sum^{k}(O_j-E_j)^2/E_j$$

où O_j et E_j nous donnent les fréquences observées et espérées respectivement.

$$\text{Or sous } H_0,\ E(X_S^2) = \sum^{k} V(O_j)/E_j = \sum^{k} \hat{d}_j * V(O_j)_*/E_j$$

$$= \sum^{k} \hat{d}_j E_j \left(\frac{n_i - E_j}{n_i}\right)/E_j$$

$$= \sum^{k} \hat{d}_j (1-P_j)$$

où $V(O_j)_*$ est la variance qui serait obtenue sous un é.a.s.a.r..
De plus par définition, si nous sommes en présence d'un é.a.s.a.r.,

$$E(X_S^2) = \sum^{k} 1(1-P_j) = k-1$$

De sorte que si l'on veut que notre nouvelle statistique ait une espérance qui soit égale à k-1, il nous faut donc la multiplier par le facteur suivant

$$F = (k-1)/\sum^{k} \hat{d}_j (1-P_j)$$

d'où le résultat

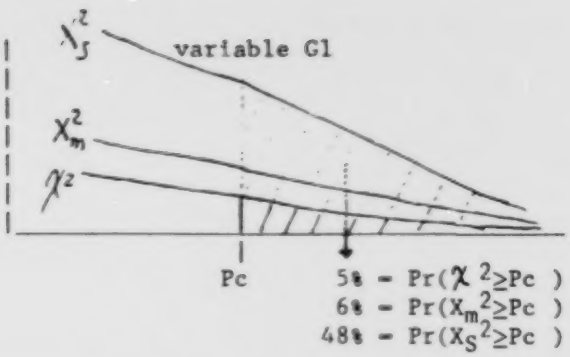
$$X_m^2 = F * X_S^2 \quad .$$

4.3 MÉCANIQUE DU TEST

Encore une fois la mécanique ici est la même que celle donnée précédemment.

4.4 AVANTAGES ET DÉSAVANTAGES

Cette méthode est plus facile à comprendre et à utiliser que l'approche précédente. On remarque également qu'en plus de bien s'adapter au plan de sondage, elle est relativement peu coûteuse à implanter. Son plus gros avantage provient du fait qu'elle ne requiert que la connaissance de la diagonale de la matrice de variance-covariance. Ce dernier point est non négligeable lorsque que l'on est en présence d'enquêtes complexes. Cependant on peut montrer que cette approche a tendance à rejeter l'hypothèse nulle H_0 un peu plus souvent que désiré comme l'illustre l'exemple suivant. L'exemple nous vient de l'enquête (GENERAL HOUSEHOLD SURVEY) [5] de laquelle 13000 ménages ont été sélectionnés. On a eu recours à la stratification lors de l'élaboration de ce plan de sondage à 3 niveaux. L'échantillonnage en est un proportionnel à la taille Pour 13 variables, G1,...G13, on a calculé le point critique correspondant au 95ième percentile en accord avec une distribution du Chi-deux à k-1 degrés de liberté. On a ensuite calculé la probabilité d'observer une valeur plus grande ou égale à ce point critique (Pc) selon la distribution de la statistique du Chi-deux standard (X_S^2) et selon la distribution de la statistique du Chi-deux modifié (X_m^2). Le graphique suivant illustre les résultats obtenus pour une de ces variables.



Le tableau II résume ces résultats pour l'ensemble des variables.

TABLEAU II. PROBABILITÉ D'OBTENIR UNE VALEUR SUPÉRIEURE OU ÉGALE A Pc SOUS LES DISTRIBUTIONS DES STATISTIQUES DU X_S^2 ET DU X_m^2

Variables	degrés liberté	X_S^2 (%)	F (facteur)	X_m^2 (%)	Chi-deux (χ^2) (%)
G1	3	48	3,27	6	5
G2	2	41	3,42	5	5
G3	3	37	2,54	6	5
G4	3	30	2,17	6	5
G5	9	16	1,23	7	5
G6	3	16	1,48	6	5
G7	6	15	1,29	6	5
G8	9	14	1,19	6	5
G9	5	10	1,14	7	5
G10	2	10	1,26	5	5
G11	4	6	1,01	6	5
G12	1	6	1,13	5	5
G13	3	6	1,02	5	5

On remarque que non seulement le facteur F joue un rôle important mais qu'il en est de même pour le nombre de degrés de liberté impliqués (la statistique du Chi-deux standard performe assez mal lorsque le nombre de degrés est élevé malgré un facteur F tout près de la valeur unitaire comme c'est le cas pour la variable G8 par exemple).

5. CONCLUSION

Étant donné les sommes d'argent et de travail considérables investies lors de l'élaboration d'enquêtes, et des implications possibles découlant de l'exploitation de ces données par de multiples utilisateurs qui émettront moult conclusions sur des sujets tout aussi nombreux que diversifiés, il apparait essentiel de s'assurer que l'échantillon sur lequel seront basées bon nombre d'analyses soit représentatif de la population dont il est issu. Nous avons présenté ici quatre techniques relativement simples permettant de détecter s'il y a eu un biais lors du mode de sélection de l'échantillon. Certaines d'entre elles sont très spécifiques au plan de sondage préconisé au recensement de la population canadienne. Par contre, il est possible de généraliser les autres techniques de façon à ce qu'elles répondent à des besoins plus particuliers. En guise de mot de la fin, j'aimerais seulement souligner l'importance de pouvoir se donner un échantillon crédible aux yeux des utilisateurs qui parfois ne sont pas toujours vendus aux vertus de l'échantillonnage.

BIBLIOGRAPHIE

[1] BANKIER, M. (1980), ''PROPOSAL FOR THE 1981 SAMPLING BIAS STUDY (DRAFT)'', STATISTIQUE CANADA

[2] PARZEN, EMANUEL, (1960), ''MODERN PROBABILITY THEORY AND ITS APPLICATIONS'', WILEY, 179.

[3] HOLT, D., SCOTT, A.J., EWINGS, P.D. (1980), ''CHI-SQUARED TESTS WITH SURVEY DATA'', STATIST. SOC., 143, 303-320

[4] RAO, J.N.K. AND SCOTT, A.J. (1981), ''THE ANALYSIS OF CATEGORICAL DATA FROM COMPLEX SAMPLE SURVEYS: CHI-SQUARED TESTS FOR GOODNESS OF FIT AND INDEPENDENCE IN TWO-WAY TABLES'', JASA, 76, 221-230

[5] OFFICE OF POPULATION, CENSUSES AND SURVEYS(1973). GENERAL HOUSEHOLD SURVEY 1973 ; ANNUAL REPORT. LONDON; H.M.S.O.

LE REMANIEMENT DU PLAN DE SONDAGE DE L'ENQUÊTE NATIONALE SUR LES FERMES DE 1988

C. JULIEN et F. MARANDA[1]

RÉSUMÉ

L'Enquête nationale sur les fermes est une enquête échantillon annuelle qui produit des estimations sur une variété de sujets reliés à l'agriculture canadienne. En 1988, on a procédé à un remaniement important du plan de sondage de cette enquête. À l'instar de l'ancien, le nouveau plan fait intervenir des bases de sondage multiples ainsi que des méthodes de stratification et d'allocation multidimensionnelles. Dans cet article, on décrit d'abord la stratégie et les méthodes employées pour le développement du nouveau plan de sondage. Ensuite, on apporte des précisions sur divers aspects du déroulement de l'enquête qui ont une influence sur la performance des nouveaux échantillons. Enfin, on évalue les résultats atteints en comparant la précision des estimations obtenues en 1988, dans un premier temps, à celle de l'enquête de 1987 et, dans un deuxième temps, à la précision espérée par le remaniement de l'enquête.

1. INTRODUCTION

L'Enquête nationale sur les fermes est une enquête échantillon probabiliste portant sur plusieurs sujets reliés à l'agriculture canadienne. Elle est menée annuellement en juin et juillet dans toutes les provinces à l'exception de Terre-Neuve où une enquête séparée est effectuée.

L'ancien plan de sondage de l'enquête datait de 1983 et il était basé sur les résultats du Recensement de l'agriculture de 1981. Or depuis 1981, la population des exploitations agricoles a subi plusieurs changements qui ont entraîné une perte d'efficacité de ce plan. De plus, les objectifs de l'enquête ont quelque peu changé au cours des ans et il était devenu impératif d'apporter des correctifs aux échantillons.

Pour ces raisons, on a procédé à un remaniement en profondeur du plan de sondage de l'enquête. Ce remaniement s'appuie sur les résultats du Recensement de l'agriculture de 1986 et il est devenu opérationnel à l'été 1988.

---

[1] C. Julien est méthodologiste, Section des méthodes d'enquêtes agricoles, Division des méthodes d'enquêtes-entreprises, Statistique Canada, Ottawa, Ontario, K1A 0T6; F. Maranda est chef, section des enquêtes agricoles, Division des méthodes d'enquêtes-entreprises, Statistique Canada, Ottawa, Ontario, K1A 0T6.

2. OBJECTIFS DE L'ENQUÊTE

L'objectif premier de l'enquête consiste à produire des estimations actuelles et fiables, reflétant les niveaux et les tendances annuelles de plus d'une centaine de variables agricoles. Ces variables se répartissent essentiellement en trois catégories: les superficies ensemencées de l'année en cours; la taille des cheptels au premier juillet; et enfin les recettes et dépenses d'exploitation pour l'année civile précédente. Ces estimations pour les catégories susmentionnées sont requises respectivement à la fin juin, en août et en septembre. En termes de précision, l'enquête vise des coefficients de variation inférieurs à 5 pourcent pour les paramètres importants à l'échelle provinciale.

Les données de l'enquête sont normalement agrégées au niveau des provinces. Cependant, surtout pour des fins analytiques, l'enquête produit également des résultats pour des régions infra-provinciales au moyen de méthodes d'estimation par le domaine.

Un autre objectif important de l'enquête est de procurer un échantillon maître duquel on choisit des sous-échantillons servant à d'autres enquêtes agricoles effectuées par Statistique Canada.

3. POPULATIONS CIBLE ET ENQUÊTÉE

La population cible comprend toutes les fermes des provinces concernées dont la vente de produits agricoles s'est chiffrée à 250 dollars ou plus au cours des 12 mois précédant le début de l'enquête. Font également partie de la population cible, les fermes qui ne satisfont pas au critère du 250 dollars en date de l'enquête mais qui anticipent réaliser au moins cette somme au cours des 12 mois suivant l'enquête. Ces dernières sont relativement peu élevées en nombre; ce sont des fermes qui ont débuté leurs activités juste avant l'enquête ou qui sont temporairement inactives.

La population enquêtée, c'est-à-dire celle qui est effectivement échantillonnée, exclut les fermes d'institutions et les fermes situées dans les réserves ou établissements indiens. Pour diverses raisons, le rapport coûts-bénéfices associé à l'obtention des données d'enquêtes pour ces deux types de fermes est très élevé; ainsi, on les exclut afin de permettre une meilleure utilisation des ressources disponibles pour la réalisation de l'enquête. La

contribution des exclusions à la production agricole nationale est faible et on l'estime en utilisant des facteurs d'ajustements calculés à partir des données du recensement.

4. BASES DE SONDAGE ET LEUR UTILISATION

La population enquêtée se répartit conceptuellement en deux groupes, le premier renfermant les fermes qui ont été dénombrées au recensement et le second toutes les autres fermes. Ces autres fermes sont celles qui ont été oubliées au recensement, ce qui constitue le sous-dénombrement, et celles dont les opérations ont commencé après le recensement, c'est-à-dire les nouvelles exploitations.

Le premier groupe est couvert, en tout ou en partie selon la province, par une ou deux bases de liste formées à même le registre des exploitations du recensement. Pour compléter les bases de liste et assurer une couverture complète de la population enquêtée, on a recours à une base aréolaire qui est créée à partir des secteurs de dénombrement agricoles. Par secteur de dénombrement, on entend la région géographique qui est assignée à un agent recenseur; de plus, un secteur est dit agricole s'il renferme au moins une ferme au recensement.

Le remaniement utilise abondamment des bases de liste car elles admettent des techniques d'échantillonnage plus efficaces et des méthodes de collecte des données moins dispendieuses qu'une base aréolaire. Le recours à une base aréolaire est cependant justifié par le besoin de pallier aux lacunes des bases de liste, notamment en ce qui concerne leur incapacité à détecter les nouvelles exploitations.

Les exigences de l'enquête en matière d'estimation et les caractéristiques de l'agriculture canadienne varient selon la région. Pour mieux tenir compte de ces variations, on divise le territoire couvert par l'enquête en trois régions et on utilise un plan de sondage distinct dans chacune d'entre elles. Les trois régions concernées sont les suivantes: les provinces des Prairies et le district de la rivière de la Paix en Colombie-Britannique; le Québec et l'Ontario; et enfin les provinces Maritimes et le reste de la Colombie-Britannique. La première de ces régions est appelée région de la Commission canadienne du blé car c'est le territoire auquel s'étend la juridiction de cet organisme.

La taille totale des échantillons dans chacune des trois régions est établie essentiellement à partir du budget global disponible pour la collecte des données. À l'intérieur de chaque région, la répartition des échantillons entre les diverses provinces et, selon le cas, entre les diverses bases, dépend de plusieurs facteurs. Les principaux facteurs en jeu sont la règle de la racine carrée de la taille de la population enquêtée, les répartitions historiques de l'enquête et les résultats de diverses analyses portant sur la précision espérée des estimations.

4.1 Région de la Commission canadienne du blé

Dans cette partie du Canada, on utilise deux bases de liste et une base aréolaire dans chaque province.

La première base de liste, notée L1, comprend essentiellement les grandes et moyennes fermes du recensement relativement à des variables clés de culture, de bétail et de dépenses. Cette liste est obtenue à l'aide d'un processus itératif qui consiste à établir un seuil pour chaque variable clé et à inclure dans la liste toutes les fermes qui excèdent au moins un de ces seuils. On ajuste indépendamment à la hausse ou à la baisse chacun des seuils de façon à ce que la liste L1 représente, une fois complétée, environ 35 pourcent des fermes de la population enquêtée et de 50 à 90 pourcent de l'activité agricole totale selon la variable clé considérée. Ces pourcentages sont retenus car l'expérience a démontré que la liste qui en résulte est stable comparativement au reste des fermes de la population enquêtée. Cette stabilité est un facteur important au développement d'un plan d'échantillonnage qui conserve son efficacité au fil des ans.

La deuxième base de liste, notée L2, renferme les fermes du recensement de plus de 20 acres qui n'ont pas été retenues dans la liste L1. L'emploi de cette base tient au fait que les données sur les cultures doivent être recueillies et compilées dans des délais très courts et la base aréolaire ne peut, pour des raisons opérationnelles, satisfaire ces délais.

La base aréolaire comprend tous les secteurs de dénombrement agricoles, sauf ceux correspondant aux réserves indiennes et aux régions où il y a peu d'activité agricole selon le recensement. Ces régions, dites marginales, comprennent surtout le nord des provinces et les zones en bordure des villes. Les rares fermes du recensement qui sont situées dans les régions marginales sont ajoutées à la liste L1.

Les bases L1 et aréolaire sont stratifiées à l'intérieur de régions infra-provinciales, nommément les types de sol, selon des variables de culture, de bétail et de dépenses. Quant à la stratification de la liste L2, elle se fait à l'intérieur des districts agricoles avec une seule variable, soit la superficie cultivée au recensement.

De chacune des deux bases de liste L1 et L2, on prélève un échantillon de fermes. De plus, un sous-ensemble, appelé noyau, de l'échantillon de liste L1 est identifié pour la collecte de renseignements plus détaillés portant sur les dépenses. La base aréolaire engendre ultimement un échantillon de segments qui sont délimités sur des cartes topographiques. Au moment de choisir ces segments, on ne connait pas l'identité des fermiers qui exploitent des terrains dans ces derniers. Cette information est obtenue par un dénombrement sur place. Suite à cela, des appariements manuels sur les noms et adresses permettent de détecter les cas de chevauchement entre les fermes de segments et l'une ou l'autre des bases de liste. Cette détection est essentielle car chaque fois que la base aréolaire est appelée à compléter une base de liste, seules les fermes de segments qui ne chevauchent pas la liste en question contribuent aux estimations. Les autres fermes de segments sont ignorées conformément au principe de l'estimation par le domaine.

Les estimations sont produites en combinant les divers échantillons de la façon suivante: listes L1 et L2, puis liste L1 et aréolaire pour les estimations respectivement préliminaires et révisées sur les cultures; liste L1 et aréolaire pour le bétail; et enfin, liste L1 noyau et aréolaire pour les dépenses. On remarque que tous les résultats font intervenir l'échantillon de liste L1. Ainsi, toutes les fermes aréolaires qui chevauchent la liste L1 ne sont jamais appelées à participer aux estimations et n'ont donc pas à compléter un questionnaire.

4.2 Québec et Ontario

Dans ces deux provinces, on a recours à une seule base de liste, appelée L1, et à une base aréolaire. La base de liste est composée de l'ensemble des fermes du recensement qui appartiennent à la population enquêtée. Quant à la base aréolaire, elle n'a pas été entièrement redéfinie, faute de ressources, à partir des résultats du dernier recensement. Ainsi, elle comprend tous les secteurs de dénombrement agricoles du Recensement de l'agriculture de 1981, à l'exception de ceux correspondant aux réserves indiennes et aux régions où l'activité agricole est faible. Ces régions, dites marginales, ont cependant été mises à jour suite au dernier recensement, ce qui a résulté en leur agrandissement comparativement à l'ancien plan de sondage de l'enquête.

Chaque base est stratifiée à l'intérieur de régions infra-provinciales selon des variables de culture, de bétail et de dépenses. De plus, lors de la stratification de la liste, on sépare au préalable les fermes incorporées, c'est-à-dire constituées en sociétés par actions, des autres et on forme indépendamment des strates dans chacun de ces deux groupes. Cette séparation préliminaire tient au fait que seules les fermes incorporées doivent rapporter leurs dépenses dans l'enquête, les dépenses des autres fermes étant obtenues à partir des dossiers fiscaux.

Des bases liste et aréolaire, on prélève respectivement un échantillon de fermes et de segments. L'échantillon aréolaire n'a pas été remanié depuis sa création en 1983, à l'exception d'un renouvellement partiel d'environ 25 pourcent qui a eu lieu la plupart des années. En 1988, un tel renouvellement n'a pas été effectué; on a simplement utilisé l'échantillon de l'année précédente duquel on a soustrait quelques segments en raison de l'agrandissement susmentionné des régions marginales.

Les résultats sont obtenus en combinant les échantillons de la façon suivante: liste, puis liste et aréolaire pour les estimations respectivement préliminaires et révisées sur les cultures; liste et aréolaire pour le bétail et également pour les dépenses des fermes incorporées. Tout comme dans la région de la Commission canadienne du blé, seules les fermes de segments qui ne chevauchent pas la base de liste sont conservées dans l'échantillon aréolaire aux fins d'estimation.

4.3 Provinces Maritimes et le reste de la Colombie-Britannique

Dans cette région, le plan de sondage ne comprend qu'une seule base de type liste appelée L1. Cette dernière consiste en l'ensemble des fermes du recensement qui font partie de la population enquêtée. Étant donné qu'une base de liste a tendance à se détériorer avec le temps et qu'il n'y a pas de base aréolaire pour la compléter, il devient alors plus difficile de couvrir entièrement la population enquêtée. Cependant, en raison du nombre relativement peu élevé de fermes dans les provinces concernées, des procédures plus poussées pour tenir à jour la liste ont été mises en place. Ces procédures permettent notamment de détecter des fermes qui ont été oubliées au recensement ou qui ont commencé leurs activités depuis lors. On estime ainsi que la base de liste assure à toute fin pratique une couverture exhaustive de la population enquêtée.

Dans chaque province, la liste est séparée en deux groupes, soit les fermes incorporées et celles qui ne le sont pas. Tout comme au Québec et en Ontario, cela tient au fait qu'on utilise les dossiers fiscaux en remplacement des données d'enquête pour les dépenses des fermes qui ne sont pas incorporées. On stratifie ensuite chacun des deux groupes selon des variables de culture, de bétail et de dépenses. De plus, au Nouveau-Brunswick seulement, la stratification est effectuée à l'intérieur de régions infra-provinciales. Une fois les strates formées, on sélectionne un échantillon de fermes qu'on utilise pour produire toutes les estimations requises.

5. MÉTHODES D'ÉCHANTILLONNAGE LISTE

Le prélèvement des échantillons de type liste s'appuie sur un plan d'échantillonnage stratifié à un degré, où les fermes constituent les unités d'échantillonnage. La stratégie et les méthodes qui sont employées pour le développement de ce plan sont essentiellement les mêmes quelles que soient la province et la liste considérées; par contre, l'agencement des méthodes et les variables clés en jeu peuvent varier d'un cas à l'autre.

La première étape consiste à identifier les fermes qui ont des caractéristiques distinctes et à procéder à un tirage complet de ces dernières. Ces fermes, dites auto-représentatives, sont essentiellement de deux types. D'abord, on retrouve celles qui ont une structure d'exploitation unique, soit les pâturages communautaires et les corporations à opérations multiples. En second lieu, il y a les fermes qui se démarquent nettement de la majorité en raison de leur très forte contribution à des variables clés de culture, de bétail et de dépenses. Le dénombrement complet de ces dernières constitue, en raison de l'asymétrie (vers la droite) des distributions traitées, une façon efficace de réduire la variance échantillonnale.

L'identification des fermes à très forte contribution se fait au moyen d'une règle dont les fondements sont intuitifs et qui a donné de bons résultats dans l'ancien plan de sondage de l'enquête. Cette règle, dite de l'écart-sigma, est appliquée indépendamment à chaque variable clé sur l'ensemble des fermes ayant une contribution non nulle pour la variable en question. Suite à cela, sont déclarées auto-représentatives toutes les fermes dont la contribution est jugée suffisamment élevée selon la règle à l'une ou l'autre des variables clés.

La règle de l'écart-sigma adaptée à l'enquête fonctionne de la façon suivante. Soit une distribution unidimensionnelle de points x_i, $i=1,\ldots,N$, $x_i > 0$ pour tout i, et soit σ son écart-type; on ordonne la distribution en ordre croissant $x_1 \leq x_2 \leq \ldots \leq x_N$; on détermine, pour la moitié de la distribution se trouvant à droite de la médiane, la distance entre chaque couple de points successifs $d_i = x_i - x_{i-1}$; soit i_o le plus petit i pour lequel $d_i \geq \sigma$, alors tous les points $i \geq i_o$ donnent lieu à des fermes auto-représentatives. Si $d_i < \sigma$ pour tout i, alors aucun point de cette distribution ne se distingue suffisamment des autres pour engendrer des fermes à tirage complet.

La seconde étape consiste à répartir le reste des fermes de la liste en strates à tirage partiel. Dans la majorité des cas, les strates sont formées à l'intérieur de régions infra-provinciales au moyen d'une procédure de classification multidimensionnelle. Cette procédure emploie simultanément neuf variables clés représentant les trois catégories usuelles: culture, bétail et dépenses. Le nombre de variables dans chaque catégorie est respectivement de un, six et deux pour un total de neuf.

Le principe sous-jacent à la stratification est le suivant: chaque ferme est un point dans un espace à neuf dimensions et on forme des ensembles de points qui sont relativement voisins en termes de distance euclidienne. Pour ce faire, on utilise des algorithmes basés notamment sur les travaux de Ward (1963) et disponibles dans les procédures FASTCLUS et CLUSTER du progiciel SAS, version 5.

Ces algorithmes sont employés successivement de la façon suivante. On utilise d'abord FASTCLUS pour regrouper les fermes concernées en 250 grappes; ensuite, on fusionne progressivement ces grappes à l'aide de CLUSTER pour finalement former les strates. On effectue une classification préliminaire avec FASTCLUS car l'emploi direct de CLUSTER avec un nombre élevé de dossiers requiert un temps d'ordinateur beaucoup trop important.

Au moment de la classification, il est impératif que chacune des trois catégories de variables contribue avec le même degré d'importance à la formation des strates. Pour ce faire, on procède à des transformations des variables initiales de stratification. Ces transformations sont exécutées de façon à ce que la somme des variables transformées dans chaque catégorie ait une moyenne de 0 et une variance déterminée, généralement 1. Pour la catégorie culture où il y a une seule variable, il suffit de la standardiser de façon usuelle en lui soustrayant sa moyenne et en divisant par son écart-type. Pour les deux autres catégories, on effectue indépendamment dans chacune d'elles deux transformations successives. Soient X_i les variables d'une catégorie donnée C.

On effectue d'abord une analyse en composantes principales pour obtenir des variables transformées Y_i. Ces nouvelles variables, de moyenne μ_i et de variance σ_i^2, sont des combinaisons linéaires des anciennes et mutuellement indépendantes. Ensuite, on standardise les Y_i pour obtenir des variables de stratification finales Z_i de la façon suivante:

$$Z_i = \frac{Y_i - \mu_i}{\left(\sum_{i \in C} \sigma_i^2\right)^{1/2}}$$

Ainsi, on constate que $\sum_{i \in C} Z_i$ possède une moyenne de 0 et une variance de 1.

Une approche empirique est employée pour décider du nombre de strates. On effectue d'abord plusieurs stratifications et allocations en faisant varier le nombre de strates. Cela fait, on construit la courbe du coefficient de variation obtenu en fonction du nombre de strates, et ce pour toutes les variables clés et bien d'autres. Ces courbes ont généralement l'allure de la Figure 1. On considère que les gains dûs à la stratification sont à toute fin pratique complètement réalisés au point où la majorités des courbes deviennent quasi horizontales. Le choix du nombre de strates est un compromis entre le point susmentionné et le désir de ne pas former trop de strates pour atténuer les effets des classifications initiales erronées et des changements de strates dans le temps, deux causes importantes d'observations aberrantes.

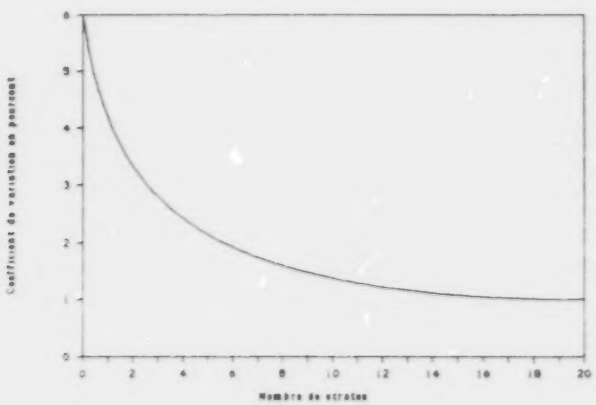


Figure 1. Courbe générale du coefficient de variation en fonction du nombre de strates

L'allocation de l'échantillon est multidimensionnelle et utilise généralement les mêmes variables clés que la stratification. L'algorithme d'allocation consiste à minimiser une combinaison linéaire des coefficients de variation au carré des variables clés considérées, sous la contrainte que la taille totale d'échantillon est fixe. En fait, soient c_i le coefficient de

variation d'une variable clé, $a_i > 0$ une constante et n_0 la taille totale d'échantillon, il s'agit de minimiser $\sum a_i c_i^2$ sous l'hypothèse que $n = n_0$. Cela fait, on procède à des ajustements de façon à ce que la taille minimale et le facteur de pondération maximal soient respectivement de 4 et 50 dans chaque strate.

Enfin, lorsque la sélection est complétée, l'échantillon est prélevé dans chaque strate de façon circulaire systématique après avoir ordonné les fermes selon leur région infra-provinciale et leurs dépenses totales d'exploitation. Pour la liste L1 de la région de la CCB, on choisit d'abord l'échantillon complet duquel on sélectionne, toujours de façon circulaire systématique, le sous-ensemble noyau.

6. MÉTHODES D'ÉCHANTILLONNAGE ARÉOLAIRE

La sélection des échantillons de type aréolaire est basée sur un plan d'échantillonnage stratifié à deux degrés. Les secteurs de dénombrement (SD) du recensement et les segments constituent respectivement les unités primaires et secondaires d'échantillonnage.

Au Québec et en Ontario, le plan d'échantillonnage aréolaire n'a pas été remanié en 1988; on en trouve une description dans Davidson (1984). Ainsi, à moins d'indication contraire, les paragraphes qui suivent décrivent la méthodologie employée dans les provinces de la région de la Commission canadienne du blé.

La première étape consiste à obtenir une mesure de l'activité agricole dans chaque SD de la base en agrégeant, au niveau du SD les données des fermes du recensement qui n'appartiennent pas à la liste L1. L'exclusion des fermes de la liste L1 des agrégations produit des distributions de SD qui réflètent fidèlement les caractéristiques des petites fermes. L'emploi subséquent des ces distributions permet de sélectionner un échantillon aréolaire qui complète la liste L1 par rapport aux petites fermes avec une efficacité accrue.

Une fois les agrégations terminées, chaque SD est traité comme une ferme pour les fins d'échantillonnage. La stratégie et les méthodes employées pour la sélection des SD sont très similaires à celles appliquées à la liste L1 de la région de la Commission canadienne du blé. En effet, on détermine d'abord des SD auto-représentatifs avec la règle de l'écart-sigma. On répartit ensuite, à l'intérieur de régions infra-provinciales, le reste des SD en strates à tirage partiel au moyen de la procédure de classification multidimensionnelle

CLUSTER du progiciel SAS. Une classification préliminaire avec FASTCLUS n'est pas nécessaire dans ce cas-ci en raison du nombre relativement peu élevé de SD à traiter. Quant aux variables clés en jeu, on utilise un ensemble quasi identique à la liste L1. Cependant les transformations appliquées aux variables clés se limitent aux standardisations usuelles. Il n'est pas nécessaire de recourir aux composantes principales car la contribution de la base aréolaire aux estimations provinciales n'est pas suffisante pour justifier une telle approche.

L'allocation aux strates est exécutée avec le même algorithme que celui de la liste et la taille minimale dans chaque strate est également fixée à 4. Cela complété, on divise la taille d'échantillon dans chaque strate par quatre et on prélève quatre répliques indépendantes de façon circulaire systématique. Le recours à des répliques facilite le calcul de la variance car il arrive souvent qu'on choisisse une seule unité secondaire par unité primaire.

Une fois les SD sélectionnés, on délimite leur contour sur des cartes topographiques et on parcellise chacun de ceux-ci en segments d'environ 7.5 kilomètres carrés ou 3 milles carrés. Ce faisant, on tente, dans la mesure du possible, d'utiliser des limites naturelles comme des routes ou rivières afin de faciliter ultérieurement le travail des interviewers sur le terrain. Puis, un échantillon aléatoire simple sans remise de segments est prélevé dans chaque SD concerné au taux minimum de un sur trente. On note cependant quelques exceptions à la règle: d'abord, on prélève des segments additionnels de façon à ce que le facteur de pondération global n'excède jamais 180; un minimum de deux segments sont choisis dans chaque SD sélectionné avec probabilité un; enfin, lorsqu'un même SD apparait dans plus d'une réplique, des dispositions sont prises pour éviter le tirage d'un même segment plus d'une fois. Toutes ces exceptions constituent cependant des cas plutôt rares.

7. RÉSULTATS DU PLAN DE SONDAGE

Le tableau 1 contient les résultats du plan de sondage des bases de liste. On y retrouve les quantités suivantes: le nombre de fermes dans la liste (N); le nombre de strates (H); la taille de l'échantillon de fermes (n); et enfin, dans les provinces où cela s'applique, le nombre de fermes dans le sous-ensemble noyau (n-noyau).

Le tableau 2 renferme les résultats du plan de sondage aréolaire dans les provinces où un tel plan est utilisé. On y retrouve les quantités suivantes:

le nombre de SD dans la base (N); le nombre de strates (H); le nombre de SD échantillonnés (n); le nombre de SD uniques échantillonnés (n-uniques); et enfin, le nombre de segments prélevés (m).

Tableau 1. Résultats du plan de sondage des bases de liste

	Liste L1				Liste L2		
Province	N	H	n	n-noyau	N	H	n
Î.-P.-É.	2830	26	451	.[c]	.	.	.
N.-É.	4273	35	550	.	.	.	.
N.-B.	3544	39	498	.	.	.	.
Québec	41380	80	6096	.	.	.	.
Ontario	72598	78	8401	.	.	.	.
Manitoba	6712	48	1364	490	18058	29	2267
Saskatchewan	15668	48	3625	1106	45798	41	4573
Alberta	13928	63	2981	909	38504	25	2973
C.-B.(Paix)[a]	494	25	190	190	1187	6	170
C.-B.(reste)[b]	17042	41	1999	.	.	.	.
Total	178469	479	26155	2695	103547	101	9983

[a] District de la rivière de la Paix en Colombie-Britannique.
[b] Colombie-Britannique, sauf le district de la rivière de la Paix.
[c] Sans objet

Tableau 2. Résultats du plan de sondage aréolaire

Province	N	H	n	n-uniques	m
Québec	2065	43	191	182	230
Ontario	2687	49	195	185	259
Manitoba	794	21	277	264	305
Saskatchewan	1496	26	328	308	477
Alberta	1623	32	328	319	434
C.-B.(Paix)[a]	54	7	36	32	58
Total	8719	178	1355	1290	1763

[a] District de la rivière de la Paix en Colombie-Britannique

8. COLLECTE DES DONNÉES

Pour les échantillons de liste, la collecte des données est effectuée en grande majorité par entrevue téléphonique. Par exception, la collecte pour le sous-ensemble noyau s'effectue par mise à la poste du questionnaire avec reprise en personne.

Pour l'échantillon aréolaire, on effectue une entrevue en personne auprès de chaque fermier qui exploite une parcelle de terre à l'intérieur du segment. Aux exploitants qui font déjà partie de la base de liste L1, on demande seulement la superficie totale de la ferme et la superficie totale en terres à bois ainsi que ces mêmes superficies situées à l'intérieur du segment. Aux exploitants qui ne font pas partie de la base de liste L1, on demande l'information susmentionnée plus les renseignements agricoles se rapportant à l'ensemble de l'exploitation.

Au bureau chef, on vérifie une fois de plus que chaque ferme de segment pour laquelle on a collecté des renseignements sur l'ensemble de l'exploitation ne chevauche pas la base de liste L1. Ceci assure que les bases L1 et aréolaire représentent des domaines mutuellement exclusifs et complémentaires. Ainsi, on peut traiter chaque base indépendamment au moment de l'estimation.

Après cette vérification, on procède à la saisie des données où toute l'information agricole est transférée sur support informatique.

9. TRAITEMENT DES DONNÉES

9.1 Vérification informatique

Une fois la saisie effectuée, la prochaine étape consiste en la vérification informatique des données. Cette vérification décèle toute entrée inacceptable, douteuse ou incohérente. La valeur d'une telle entrée peut être exacte mais peut aussi provenir d'une erreur de saisie ou d'une réponse incomplète ou erronée de la part de l'exploitant.

Une entrée inacceptable rend tout traitement subséquent de l'enregistrement inutile. L'erreur exige une correction manuelle, c'est-à-dire qu'une personne doit décider de la correction appropriée et l'effectuer sur l'enregistrement. Par exemple, lorsqu'un répondant omet de déclarer l'unité de superficie avec laquelle il rapporte l'utilisation de ses terres, une personne

doit, en analysant les superficies rapportées ou en consultant les données du recensement, déterminer si l'exploitant a déclaré des acres, des hectares ou des arpents.

Une entrée douteuse est décelée lorsqu'une valeur, en termes absolus ou par rapport à d'autres valeurs sur l'enregistrement, est anormalement élevée ou basse. Si la valeur est incorrecte alors elle est corrigée manuellement. Voici quelques exemples. Une erreur de saisie où 100 000 000 dollars en salaires versés apparait au lieu du 1 000 000 dollars rapporté. Un apiculteur qui rapporte ses abeilles comme étant 1 000 000 à la variable autres têtes de bétail sera corrigé à 0 car ces exploitations ne font pas partie de la population cible. Enfin, l'enregistrement d'un exploitant qui déclare 10 acres de terre avec 10 000 porcs sera corrigé lorsqu'un analyste en la matière le justifie.

Une entrée incohérente est telle que sa valeur correcte peut être déterminée de façon non-aléatoire par d'autres valeurs sur le même enregistrement. L'entrée erronée est corrigée automatiquement par imputation déterministe. Par exemple, un répondant qui déclare divers types de porcs mais qui omet le nombre total de têtes se verra imputer automatiquement la somme des divers types au total. Par contre, l'entrée peut être corrigée manuellement lorsqu'il s'agit d'une erreur de saisie.

Il arrive qu'une ou plusieurs variables n'ont pas de valeur et qu'elles ne peuvent être corrigées par imputation déterministe. Dans cette situation, dite de non-réponse partielle, on procède au dernier traitement, soit l'imputation des valeurs manquantes.

9.2 Imputation des valeurs manquantes

L'imputation pour les valeurs manquantes se fait par donneur, c'est-à-dire en recourant à une méthode d'imputation qui fait appel aux données d'un autre enregistrement. Pour ce faire, on apparie l'enregistrement incomplet, le receveur, au donneur acceptable qui lui est le plus près selon la valeur de quelques variables disponibles chez le receveur.

Une fois le donneur acceptable le plus près identifié, le transfert des données se fait la plupart du temps soit en copiant directement les données de l'enregistrement donneur à l'enregistrement receveur, soit en copiant la répartition indiquée des valeurs du donneur au receveur.

Finalement, la non-réponse totale provenant du refus de répondre ou de l'impossibilité de contacter l'exploitant, ne permet pas d'imputation par donneur le plus près, faute de variables d'appariement. Ces cas de non-réponse sont traités lors de l'estimation.

10. ESTIMATION

10.1 Bases de liste

Pour la base de liste, les estimateurs employés sont ceux d'un échantillon aléatoire simple stratifié, c'est-à-dire;

$$\hat{Y}_h = \frac{N_h}{n_h} \sum_{i=1}^{n_h} y_{hi}$$

$$var(\hat{Y}_h) = N_h(N_h - n_h)\frac{s_h^2}{n_h}$$

$$\text{où} \quad s_h^2 = \frac{\sum_{i=1}^{n_h} y_{hi}^2 - \frac{\left(\sum_{i=1}^{n_h} y_{hi}\right)^2}{n_h}}{n_h - 1}$$

où N_h est le nombre de fermes dans la strate h et n_h est le nombre de fermes desquelles on a obtenu une réponse, complète ou partielle, dans la strate h. À noter que ce nombre de fermes est inférieur à la taille initiale de l'échantillon; ainsi on compense la non-réponse totale en ajustant à la hausse le facteur de dilatation. Ceci équivaut à imputer la moyenne de la strate aux valeurs manquantes et à utiliser le facteur de dilatation original. De plus, cet ajustement engendre une augmentation de la variance des estimations.

10.2 Base aréolaire

Suite au plan d'échantillonnage décrit à la section 6, les estimateurs sont évalués à l'intérieur de chaque réplique.

L'information agricole se rapportant à toute la ferme de segment est recueillie pour faciliter la collecte des données; mais, pour la base aréolaire, on veut seulement la contribution de cette information se rapportant à l'activité agricole à l'intérieur du segment. Pour ce faire, on réduit les valeurs observées pour chaque ferme par un facteur w; ce facteur est la superficie de la ferme, excluant les terres à bois, située à l'intérieur du segment, divisée par la superficie totale de la ferme excluant toutes les terres à bois. Par la suite, on obtient une mesure de l'activité agricole dans le segment en additionnant les valeurs réduites pour chaque ferme à l'intérieur du segment.

$$\hat{Y}_{hij}^{(r)} = \sum_{k} w_{hijk}^{(r)} y_{hijk}^{(r)}$$

où $w_{hijk}^{(r)} = \dfrac{\text{(superf. de la ferme ds le segment - terres à bois ds le segment)}}{\text{(superficie totale de la ferme - terres à bois)}}$

est la réduction apportée à la ferme k située dans le segment j, SD i, réplique r et strate h.

Le segment étant notre unité d'échantillonnage ultime, on peut procéder à l'estimation en employant la méthode usuelle pour un plan de sondage à deux degrés, c'est-à-dire;

$$\hat{Y}_{h}^{(r)} = \frac{N_h}{n_h} \sum_{i=1}^{n_h} \frac{M_{hi}}{m_{hi}} \sum_{j=1} \hat{Y}_{hij}^{(r)}$$

$$\hat{Y}_{h} = \frac{\sum_{r=1}^{R} \hat{Y}_{h}^{(r)}}{R}$$

$$var(\hat{Y}_{h}) = \frac{1}{(R-1)R} \left(\sum_{r=1}^{R} (\hat{Y}_{h}^{(r)})^2 - \frac{\left(\sum_{r=1}^{R} \hat{Y}_{h}^{(r)} \right)^2}{R} \right)$$

où N_h est le nombre de SD dans la strate h;
n_h est le nombre de SD dans la réplique r de la strate h;

M_{hi} est le nombre de segments dans le SD hi de la réplique r;
m_{hi} est le nombre de segments dans l'échantillon du SD hi de la réplique r; et R est le nombre de répliques.

Enfin, les estimations du niveau provincial s'obtiennent en additionnant les estimations provenant de la base de liste et de la base aréolaire. Il en est de même pour la variance des estimations puisque les deux échantillons sont indépendants.

11. ÉVALUATION DES RÉSULTATS DE L'ENQUÊTE

Pour évaluer les résultats atteints lors de l'enquête de 1988, la précision des estimations obtenues est comparée, dans un premier temps, à celle de l'enquête de 1987 et, dans un deuxième temps, à la précision espérée par le remaniement de l'enquête.

11.1 Enquête de 1988 contre enquête de 1987

Deux tendances s'opposent lorsque l'on compare la précision des estimations de 1988 à celle des estimations de 1987. D'un côté, les estimations de 1988 devraient être plus précises puisque le plan de sondage de 1987 était vieux de 4 ans déjà. Par contre, deux raisons militent en faveur d'une précision plus élevée des estimations de 1987. Premièrement, la technique pour traiter la non-réponse totale employée en 1987 remplaçait l'enregistrement de la ferme non déclarée par celui, complet, d'une autre ferme de la même strate. De plus, le nouveau plan de sondage présente une réduction de 10 pourcent de la taille de l'échantillon de liste L1 dans les provinces de la région de la Commission canadienne du blé.

La précision est comparée en employant le coefficient de variation (CV) des estimations du niveau provincial provenant de l'union des bases de liste L1 et aréolaire. Ces estimations sont celles de plusieurs variables clés dont le CV en 1987 n'excède pas 20%.

La comparaison de la précision de 234 estimations est présentée à l'aide de graphiques à l'annexe A. Sur ces graphiques, chaque carré représente le CV d'une estimation tel qu'atteint en 1987, en abcisse, et atteint en 1988, en ordonnée. De plus, on présente la fréquence (en pourcentage) des carrés situés à l'intérieur de chaque zone délimitée par les droites Y=X/2, Y=X et Y=2X.

On observe que près de 60% des estimations des cultures sont plus précises en 1988. Celles qui le sont moins, le sont par peu dans la plupart des cas. Pour le bétail, plus de 90% des estimations sont plus précises en 1988. De plus, 32% de celles-ci sont même deux fois plus précises. Enfin, plus de 60% des estimations des dépenses sont plus précises en 1988. Par contre, certaines de ces estimations sont beaucoup moins précises, même plus que deux fois moins précises. Ces estimations proviennent du Québec et de l'Ontario où on collecte les dépenses d'exploitation seulement chez les fermes constituées en sociétés par actions, c'est-à-dire incorporées. Or, dans ces provinces, la forme juridique d'une ferme est difficile à définir, autant lors du recensement que lors de l'enquête.

On en conclut que, malgré la diminution de la taille de l'échantillon planifiée, lors du remaniement de l'enquête, et observée, lors du traitement pour la non-réponse totale, l'enquête de 1988 a fourni des estimations plus précises pour chaque catégorie de variables.

11.2 Précision atteinte contre précision espérée

On s'attend à ce que la précision atteinte soit moins bonne que la précision espérée, et ce, pour deux raisons. Premièrement, l'ajustement des facteurs de dilatation pour compenser la non-réponse totale entraine une augmentation de la variance. Or, on a observé un taux de non-réponse totale variant entre 2 et 13% selon la province. Au niveau du Canada ce taux s'établit à 10 pourcent. Le tableau 3 donne les détails des taux provinciaux de non-réponse.

Tableau 3. Taux de non-réponse par province

Province	Refus	Non-Contact	Total
Î.-P.-É.	0.00%	3.55%	3.55%
N.-É.	0.00%	2.18%	2.18%
N.-B.	0.00%	1.61%	1.61%
Québec	1.71%	6.56%	8.27%
Ontario	2.27%	11.11%	13.38%
Manitoba	3.45%	4.03%	7.48%
Saskatchewan	4.06%	6.46%	10.52%
Alberta	2.68%	7.95%	10.63%
C.-B.	1.78%	10.28%	12.06%
Total	2.32%	8.11%	10.43%

Deuxièmement, les données qui ont servi à créer la base de sondage proviennent du recensement agricole de 1986. D'une part, ces données sont sujettes à des erreurs, et, d'autre part, il s'effectue une certaine détérioration de la base de sondage dans le temps, suite aux changements dans l'activité agricole.

La précision est comparée en employant le coefficient de variation des estimations du niveau provincial provenant de la base de liste L1 seulement. Ces estimations sont celles de plusieurs variables clés dont le CV espéré n'excède pas 20%.

La comparaison de la précision de 288 estimations est présentée à l'aide de graphiques à l'annexe B. Sur ces graphiques, chaque carré représente, pour une estimation, le CV espéré, en abcisse, et atteint en 1988, en ordonnée. De plus, on présente la fréquence (en pourcentage) des carrés situés à l'intérieur de chaque zone délimitée par les droites Y=X, Y=2X et Y=3X.

Pour les cultures et le bétail, environ 90% des estimations ont une précision acceptable, compte tenu du taux de non-réponse, puisque le rapport (CV atteint / CV espéré) se situe près de 1.0 et 1.5. Pour les dépenses, on distingue deux tendances. D'abord, fait remarquable, le CV atteint est inférieur dans 28% des cas à celui espéré; ceux-ci proviennent en grande majorité de la région de la Commission canadienne du blé. Par contre, on remarque plusieurs cas où la précision atteinte est faible; ces cas proviennent du Québec et de l'Ontario pour les raisons mentionnées à la section 11.1.

Enfin, on a effectué une autre étude où on a comparé la précision atteinte à la précision espérée en ajustant cette dernière pour la taille de l'échantillon effectivement observée. On a constaté que la fréquence des estimations plus que deux fois moins précises que prévu passait de 12 à 5% pour les cultures, de 9 à 5% pour le bétail et de 31 à 7% pour les dépenses.

On en conclut qu'en général la précision atteinte est acceptable et diffère de la précision espérée principalement à cause du traitement pour la non-réponse totale. Ceci indique que la plan de sondage est robuste et que la base de sondage de liste L1 est adéquate. Par contre, on a obtenu des estimations moins précises pour les dépenses à cause d'un problème dans l'identification des fermes incorporées au Québec et en Ontario lors du recensement et de l'enquête. Enfin, on a noté une certaine détérioration de la base de sondage, vieille de deux ans lors de l'enquête, reflétée par les faillites et les ventes de ferme.

12. CONCLUSION

De façon générale, les résultats de l'enquête ont été sensiblement améliorés suite au remaniement du plan de sondage. Également, la réduction des tailles échantillonnales a permis de réaliser des économies et de réduire de façon appréciable le fardeau de réponse. Cependant, certaines difficultés persistent, surtout au niveau des dépenses des fermes incorporées au Québec et en Ontario. Des travaux additionnels sont nécessaires pour résoudre ces difficultés.

BIBLIOGRAPHIE

Davidson, G. (1984), "Enquête nationale sur les fermes de 1983: Exposé du plan d'échantillonnage et des méthodes d'estimation", document de travail interne de la Division des méthodes d'enquêtes-entreprises, Section des enquêtes agricoles.

Cochran, W. G. (1977), "Sampling techniques" 3rd edition, Wiley, New York.

Germain, M.-F., Dolson, D. et Maranda, F. (1989), "Le remaniement du plan de sondage de l'enquête nationale sur les fermes de 1988", document de travail interne de la Division des méthodes d'enquêtes-entreprises, Section des enquêtes agricoles.

Lynch, Jenny (1988), "Cas spéciaux d'estimation dans l'enquête nationale des fermes de 1988", document de travail interne de la Division des méthodes d'enquêtes-entreprises, Section des enquêtes agricoles.

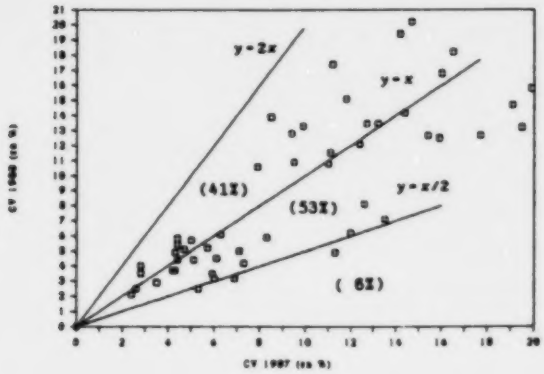
SAS Institute Inc. SAS User's Guide: Statistics, Version 5 Edition. Cary, NC: SAS Institute Inc., 1985

Ward, J.H. (1963), "Hierarchical Grouping to Optimize an Objective Function", Journal of the American Statistical Association, 58, 236-244.

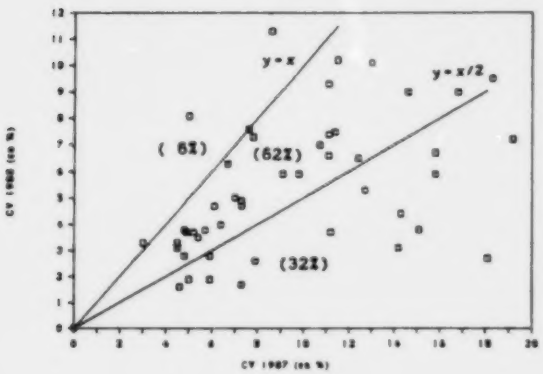
ANNEXE A

Comparaison de la précision des estimations de l'enquête de 1988 à celle de 1987.

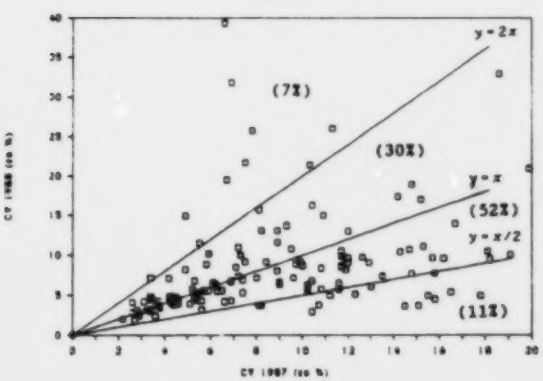
Cultures



Bétail



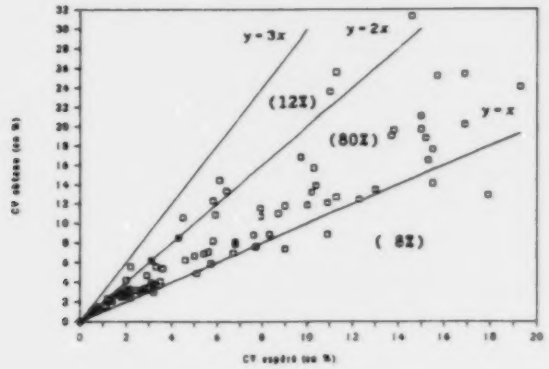
Dépenses



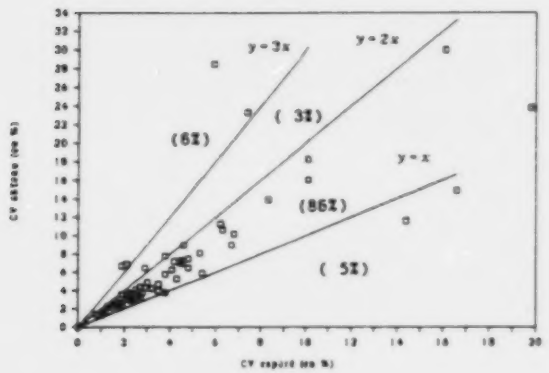
ANNEXE B

Comparaison de la précision des estimations de l'enquête de 1988 à celle espérée par le remaniement de l'enquête.

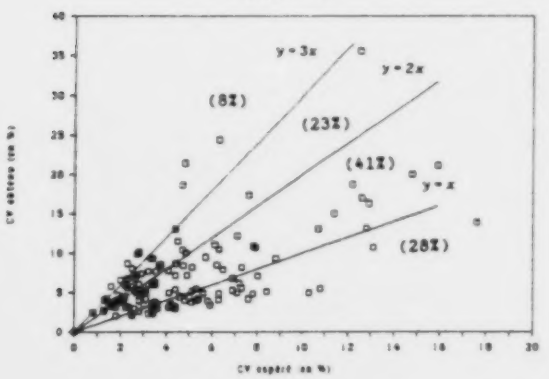
Cultures



Bétail



Dépenses



ESTIMATION DE LA MOYENNE ET DE LA VARIANCE D'ÉCHANTILLONNAGE POUR UN PLAN QUELCONQUE A L'AIDE DU JACKKNIFE

SYLVAIN MASSE[1]

Résumé

Soit n observations $\{y_i, i=1,2,..n\}$ et $\{w_i, i=1,2,..n\}$ les poids associés à ces observations selon la méthode d'échantillonnage utilisée. L'estimateur $y_p=\Sigma w_i y_i/\Sigma w_i$ est asymptotiquement non-biaisé et nous montrons à l'aide de la théorie du jackknife que sa variance peut être estimée par $var(y_p)= \Sigma(w_i/w)^2(y_i-y_p)^2/(n(n-1))$ quand le plan est à probabilités variables. Si l'échantillonnage est stratifié ou s'il y a post-stratification, il faudra soustraire des observations l'estimation de la moyenne des strates avant d'utiliser cette dernière formule. Ces résultats sont appliqués à l'enquête Som-R. Dans cette enquête, le tirage est fait à probabilités variables et l'on post-stratifie par rapport à l'âge et au sexe. Par simulation nous montrons que l'intervalle $y_p \pm 1.96(var(y_p))^{.5}$ est un intervalle de confiance de seuil 95% quand l'échantillonnage est stratifié, avec post-stratification et même plus complexe. Les lois simulées sont deux lois de Bernoulli, deux normales, une log-normale et une khi-carré.

INTRODUCTION

Plusieurs articles ont été écrits au cours des dernières années en vue d'estimer la variance d'échantillonnage suite à diverses méthodes d'échantillonnage complexe. Mais les formules produites dans ces articles pour estimer la variance d'échantillonnage sont souvent fort complexes et ainsi fort peu utilisées car elles nécessitent une grande compétence statistique et beaucoup de temps pour être employées. En pratique, la post-stratification est souvent utilisée pour corriger la non-représentativité de l'échantillon même si l'on connaît peu son impact sur la variance d'échantillonnage surtout lorsqu'elle est appliquée suite à un plan complexe. C'est pourquoi beaucoup de firmes spécialisées dans le domaine de sondage utilisent la formule $1/n^{0.5}$ (où n est la taille de l'échantillon) à toutes les sauces pour estimer la marge d'erreur d'échantillonnage, même si celle-ci ne devrait être appliquée que dans le cas d'un échantillonnage aléatoire simple.

L'approche étudiée, dite technique de pondération, est simple et consiste pour un plan quelconque à identifier d'abord une pondération, i.e. à choisir des poids pour chacune des observations. La pondération définie sert par la suite à construire un estimateur pondéré qui permet d'estimer la moyenne des unités d'une population. La variance d'échantillonnage de l'estimateur pondéré est par la suite estimée par une formule simplifiée extraite de la théorie du jackknife. Cette approche fera l'objet d'une discussion dans le troisième chapitre.

Le but ultime est d'appliquer cette théorie à l'enquête téléphonique Som-R qui est une enquête téléphonique provinciale complexe dans les régions du Québec métropolitain, du Montréal métropolitain et dans le reste de la province. Dans cette enquête, le tirage est fait à probabilités variables et l'on post-stratifie par rapport à l'âge et au sexe.

---

1. Sylvain Masse est statisticien à la maison de recherche SOM.

On supposera que la cueillette est faite avec remise et que le tirage est fait dans une population infinie. A notre avis, cette procédure n'apporte pas une perte substantielle dans l'estimation. L'approche est globale et ne se résume pas à l'obtention d'estimateurs exclusifs au Som-R. La théorie énoncée permettra d'étendre les fruits de cette recherche à d'autres plans d'échantillonnage.

Dans le premier chapitre de ce texte, le plan d'échantillonnage du Som-R sera énoncé. Le second chapitre portera sur l'application de la théorie classique pour estimer des moyennes et les variances d'échantillonnage suite à de la post-stratification appliquée à de l'échantillonnage à probabilités variables. Dans le troisième chapitre, l'estimateur pondéré et la théorie du jackknife seront introduits. Dans ce même chapitre, les nouveaux estimateurs obtenus feront l'objet d'une réflexion dans le cas de l'échantillonnage stratifié et à probabilités variables. Dans le quatrième chapitre, un intervalle de confiance déduit du troisième chapitre sera comparé par la présentation des résultats de simulations par la méthode de Monte Carlo dans le cadre d'échantillonnages stratifiés non-proportionnellement et post-stratifiés. Finalement, le dernier chapitre va consister d'abord à présenter des estimateurs de la moyenne et de la variance d'échantillonnage pour le Som-R issus de la théorie classique et de la technique de pondération. Ce chapitre se terminera par la présentation des résultats de simulations de Monte Carlo dans le cas d'une imitation du plan d'échantillonnage du Som-R.

CONVENTION UTILISÉE

Tout au long de ce document, le caractère gras sera utilisé pour signifier la moyenne et les majuscules seront associées aux unités d'une population tandis que les minuscules seront associées aux unités d'un échantillon.

Le symbole Σ' sera utilisé pour représenter la somme des éléments d'une population tandis que Σ représentera la somme des éléments d'un échantillon. Les bornes des sommes — la valeur de départ et la valeur finale— seront fournies à droite des expressions. Par exemple,

$$\Sigma\Sigma\Sigma \text{ (expression)} \qquad (r=1,...,3)(h=1,...6)(i=1,...,n_{rh})$$

signifie que, dans la première somme, r varie de 1 à 3, dans la deuxième somme, h varie de 1 à 6 et, dans la dernière somme, i varie de1 à n_{rh}.

Par exemple, $\mathbf{Y}= \Sigma' Y_i / N$ (i=1,..N) sera la moyenne des N unités de la population et Y=NY, le total des éléments de cette population. Par contre, $\mathbf{y}=\Sigma y_i/n$ (i=1,2,...,n) sera la moyenne des n unités de l'échantillon.

1. LE SOM-R

1.1 Le plan d'échantillonnage du Som-R

Un sondage omnibus Som-R est produit mensuellement au niveau provincial par la compagnie Som. On entend par sondage omnibus, une enquête qui porte sur plusieurs sujets. La cueillette de cette enquête est effectuée par téléphone auprès de la population adulte du Québec. L'échantillonnage de cette enquête est complexe. Chaque ménage est sélectionné au hasard à l'intérieur de trois strates et un seul répondant par ménage est admissible.

L'échantillonnage utilisé dans cette enquête est stratifié et comporte deux degrés. Le premier degré du plan consiste à faire la sélection de ménages dans trois strates de la province de Québec. Ces trois strates de la province de Québec sont : la région métropolitaine de Québec, la région métropolitaine de Montréal et le reste de la province. Les unités primaires d'échantillonnage sont obtenues à l'aide du logiciel CONTACT. Ce logiciel permet la sortie aléatoire informatique de numéros de téléphone dans ces trois strates. Habituellement, le nombre de numéros de téléphone sortis dans chaque strate n'est pas calculé proportionnellement à la population de chacune des strates mais fixé en fonction de la précision désirée dans chaque strate. Jusqu'à cinq appels sont faits pour rejoindre les ménages choisis

par CONTACT. La compagnie Som a habituellement pour objectif de compléter 300 entrevues auprès d'adultes du Québec métropolitain, 400 entrevues auprès d'adultes du Montréal métropolitain et 400 entrevues dans le reste de la province.

La deuxième degré consiste à sélectionner une personne de dix-huit ans ou plus à l'intérieur des ménages. En vue de s'assurer de la sélection aléatoire du répondant, celui-ci est choisi à l'aide d'une méthode issue de celle de Kish(1949). Cette méthode décrite dans un article de Landry(1987) a été présentée au colloque sur les applications de la statistique dans le cadre du 55 ième congrès de l'ACFAS. Selon la méthode utilisée, l'enquêteur établit une liste des personnes admissibles dans le ménage et les numérote du plus jeune au plus vieux. Une grille choisie parmi 12 est pré-imprimée sur le questionnaire et indique à l'enquêteur le rang de l'adulte à choisir selon le nombre de répondants admissibles.

L'échantillon du Som-R ne peut être considéré représentatif de la population visée. Par exemple, une plus faible participation ou un taux de refus plus élevé est observé chez certains groupes socio-économiques. Les hommes sont plus difficiles à atteindre. Les adultes sont plus difficiles à rejoindre s'ils sont jeunes. Certains groupes sont donc sous-représentés dans l'échantillon. Cette non-représentativité sera corrigée par la post-stratification.

2. LA POST-STRATIFICATION

La post-stratification tend à imiter la stratification et est un moyen de produire un ajustement pour la non-réponse. C'est la principale méthode utilisée pour tenir compte de la non-réponse. L'emploi de la post-stratification peut conduire à une réduction de la variance d'échantillonnage de plusieurs estimés. Mais, elle est surtout utilisée pour réduire le biais qui est occasionné par une moins bonne couverture de certains segments de la population. Le lecteur intéressé pourra consulter, pour de plus amples détails, le rapport du groupe d'études sur les données incomplètes du comité des statistiques nationales du National Research Council (Panel On Incomplete Data (1983)).

2.1 Suite à un échantillonnage simple

Suite à un tirage aléatoire simple avec remise d'un échantillon de taille n, les unités peuvent être classées dans h strates d'une population dont les tailles N_h sont connues. L'estimateur usuel de la moyenne est alors remplacé par l'estimateur $y_{pst} = \Sigma\, W_h\, y_h$ (h=1,2,...,L) où $W_h=N_h/N$ et y_h est la moyenne des n_h observations de la strate h. L'estimateur post-stratifié y_{pst} est un estimateur non-biaisé de la moyenne **Y**.

Dans une démonstration de Cochran (1977,p134-135), cette méthode est dite aussi précise que l'échantillonnage stratifié proportionnel pour estimer la moyenne pourvu qu'on obtienne au moins 20 observations dans chaque strate et que les erreurs sur les poids W_h puissent être ignorées.

En se basant sur des résultats de Stephan (1945,p50-51), Cochran a montré que la variance de l'estimateur pouvait être approchée par

$$\text{Var } y_{pst} \approx (1/n) \Sigma\, W_h\, S^2_h + (1/n^2) \Sigma\, (1-W_h)\, S^2_h \qquad (h=1,2,...,L)$$
$$\text{où } S^2_h = \Sigma' (Y_{hi} - Y_h)^2 / N_h \qquad (i=1,2,...,N_h)$$

$$\text{Ainsi, } \quad \text{var } y_{pst} = (1/n) \Sigma\, W_h\, s^2_h + (1/n^2) \Sigma\, (1-W_h)\, s^2_h$$
$$\text{où } s^2_h = \Sigma\, (y_{hi} - y_h)^2 / (n_h - 1) \qquad (i=1,2,...,n_h)$$

sera l'estimateur qui pourra être utilisé pour estimer la variance de l'estimateur post-stratifié suite à un tirage aléatoire de n unités.

2.2 Suite à un tirage aléatoire à probabilités variables

2.2.1 Introduction à l'échantillonnage à probabilités variables

Dans l'échantillonnage aléatoire simple, les unités ont toutes la même chance d'être choisies. Ce n'est pas le cas d'un échantillonnage qui utilise la sélection d'un seul individu par ménage. La probabilité qu'un adulte soit choisi à l'intérieur de son ménage sera inversement proportionnelle au nombre d'adultes du ménage. C'est pourquoi, on parle d'échantillonnage à probabilités variables.

Le tirage sans remise est plus compliqué que celui avec remise et ne sera pas discuté. La population étant supposée suffisamment grande, nous considérons que les estimateurs utilisés pour un tirage à probabilités variables avec remise pourront aussi être utilisés même si le tirage est sans remise. Le lecteur intéressé pourra consulter le livre de Som (1976,61-74) pour certains détails.

Considérons un échantillon de taille n tiré avec remise d'une population de taille N où chaque unité i a une probabilité connue π_i d'être tirée. Soit y_i (i=1 ...n), les valeurs de la variable. L'estimateur non-biaisé de la moyenne Y des valeurs s'exprime alors comme:

$$y_{pv1} = (1/nN) \sum y_i / \pi_i \qquad (i=1,2,...,n)$$

La variance de cet estimateur est $\text{Var } y_{pv1} = \sum' \pi_i ((Y_i/\pi_i) - NY)^2 /nN^2$ (i=1,2,...,N)
et l'estimateur non-biaisé de cette variance est

$$\text{var } y_{pv1} = \sum ((y_i/\pi_i) - Ny_{pv1})^2/n(n-1)N^2 \quad (i=1,2,...,n)$$

La démonstration de ces résultats est présentée dans Cochran (1977,p253-254).

La sélection d'un adulte par ménage avec probabilité inverse au nombre d'adultes du ménage est un exemple d'échantillonnage à probabilités variables. D'une population de taille N de M ménages, on tire d'abord m ménages avec remise. Par la suite, on sélectionne un adulte par ménage. Soit a_i (i=1,2,...,m), le nombre d'adultes de chaque ménage échantillonné. La probabilité qu'un adulte soit choisi à un tirage donné est alors l'inverse du produit de a_i par M ($\pi_i = 1/a_iM$) . Soit A, le rapport de N sur M, le nombre moyen d'adultes par ménage. L'estimateur de la moyenne d'une variable y est dans cet exemple

$$y_{m1} = (M/mN) \sum a_i y_i = (1/Am) \sum a_i y_i \qquad (i=1,2,...,m)$$

et celui de la variance d'échantillonnage de y_{m1} est

$$\text{var } y_{m1} = \sum [(a_iy_i/A) - y_{m1}]^2/ m(m-1) \quad (i=1,2,...,m)$$

2.2.2 Déduction des estimateurs post-stratifiés

Il est possible de déduire un estimateur de la variance de l'estimateur post-stratifié pour un tirage à probabilités variables et quelques autres tirages qui peuvent être considérés à probabilités variables. Pour ce faire, nous considérons un tirage aléatoire avec remise d'unités dont la probabilité de sélection est notée π_{hi} pour l'unité i de la strate h. L'appartenance à une strate h d'une unité est connue suite au tirage et la valeur d'une observation appartenant à la strate h est notée y_{hi}. Le nombre d'unités obtenues dans chacune des strates de taille N_h est noté n_h et $W_h=N_h/N$. Dans ce cas, il est recommandé qu'il y ait au moins 20 unités dans chacune des strates pour pouvoir utiliser les prochains résultats sinon il faudra faire des regroupements.

Les résultats précédents sur le tirage à probabilités variables peuvent être appliqués au niveau de chaque strate conditionnellement à la connaissance des n_h. Toutefois la notation est modifiée. Ici, la moyenne Y_h des unités Y_{hi} de la strate h est estimée par l'estimateur sans biais y_h qui est défini par

$$y_h = (1/N_h n_h) \Sigma\, y_{hi} / \pi'_{hi} \qquad (i=1,2,...,n_h) \qquad (2.1)$$

où $\pi'_{hi} = \pi_{hi}/P_h$ = Prob(choisir i | i est dans la strate h)
et $P_h = \Sigma' \pi_{hi}$ $(i=1,2,...,N_h)$.

La variance conditionnelle de cet estimé est

$$\text{Var } y_h = \Sigma' \pi'_{hi}((Y_{hi}/\pi'_{hi}) - N_h Y_h)^2 / n_h N_h^2 \qquad (i=1,2,...,N_h)$$

L'estimateur non-biaisé de la variance de y_h est donc

$$\text{var } y_h = \Sigma\, ((y_{hi}/\pi'_{hi}) - N_h y_h)^2 / n_h(n_h-1) N_h^2 \qquad (i=1,2,...,n_h)$$

Puisque l'estimateur y_h est non-biaisé, $y_{pst} = \Sigma W_h y_h$ (h=1,2,...,L) défini à la section 3.1 sera aussi non-biaisé comme lors du tirage aléatoire simple. Il reste à évaluer la variance de y_{pst} en utilisant les variances conditionnelles. Commençons d'abord par évaluer la variance de y_h.

$$\begin{aligned}
\text{Var } (y_h) &= E\,[\text{Var } (y_h \mid n_h)] + \text{Var } E\,[y_h | n_h] \\
&= E\, \Sigma' \pi'_{hi}((Y_{hi}/\pi'_{hi}) - N_h Y_h)^2 / n_h N_h^2 \\
&\qquad \text{vu que } E\,[y_h | n_h] \text{ ne dépend pas de } n_h \\
&= [\Sigma' \pi'_{hi}((Y_{hi}/\pi'_{hi}) - N_h Y_h)^2 / N_h^2]\; E\,(1/n_h) \quad (i=1,2,...,N_h) \\
&= \text{Var } (Y_h)\; E\,(1/n_h)
\end{aligned}$$

Or $E[1/n_h] \approx (1/nP_h) + (1-P_h)/(n^2 P_h^2)$ (développement en série de Taylor)

Ainsi

$$\begin{aligned}
\text{Var } y_{pst} &= \Sigma W_h^2 \text{ Var } (y_h) \qquad (h=1,2,...L) \\
&\approx \Sigma W_h^2 [(1/nP_h) + (1-P_h)/(n^2 P_h^2)] \text{ Var } Y_h \qquad (2.2) \\
&\approx \Sigma W_h 2 [(1/nP_h) + (1-P_h)/(n^2 P_h^2)] \times \\
&\quad [\Sigma' \pi'_{hi}((Y_{hi}/\pi'_{hi}) - N_h Y_h)^2 / N_h^2]
\end{aligned}$$

Donc en remplaçant Var Y_h par son estimateur var Y_h dans l'équation 2.2, on obtient l'estimateur cherché de la variance post-stratifiée:

$$\begin{aligned}
\text{var } y_{pst} &= \Sigma W_h^2 [(1/nP_h) + (1-P_h)/(n^2 P_h^2)] \text{ var } Y_h \quad (h=1,..,L) \qquad (2.3) \\
\text{où var } Y_h &= \Sigma\, ((y_{hi}/N_h \pi'_{hi}) - y_h)^2 / (n_h - 1) \qquad (i=1,2,...,n_h)
\end{aligned}$$

Nous allons maintenant utiliser cette théorie pour estimer la variance de l'estimateur post-stratifié suite au tirage d'un seul adulte par ménage obtenu suite à un échantillonnage aléatoire simple avec remise. Les adultes des ménages recueillis sont regroupés dans L strates dont on connaît les tailles dans la population. En faisant référence au cas que nous avons vu au début de cette section, la probabilité qu'un individu i de la strate h soit choisi est

$\pi_{hi}=1/MA_{hi}$ où A_{hi} est le nombre d'adultes dans le ménage de la i ème personne de la strate h et M, le nombre de ménages dans la population. Il reste à évaluer la probabilité conditionnelle qui va être utile à l'utilisation de la formule 2.1.Cette dernière pour la strate h serait:

$$\pi'_{hi} = \pi_{hi} / \Sigma' \pi_{hj} \qquad (j=1,2,...,N_h)$$
$$= (1/A_{hi}) / \Sigma' (1/A_{hj}) \qquad (j=1,2,...,N_h)$$

L'estimateur post-stratifié de la moyenne est donc

$$y_{mp} = \Sigma W_h y_{mh} \qquad (h=1,2,...,L)$$
où $$y_{mh} = (\Sigma'(1/A_{hj}) /m_h N_h) \Sigma a_{hi} y_{hi} \qquad (j=1,2,...,N_h)$$
$$= (1/A)_h \Sigma a_{hi} y_{hi} /m_h \qquad (i=1,2,...,m_h)$$

m_h ,le nombre de ménages parmi m dont l'adulte choisi est classé dans la strate h
a_{hi} ,le nombre d'adultes dans le i ème ménage de l'échantillon de la strate h
y_{hi} ,la caractéristique mesurée du i ème ménage de l'échantillon de la strate h
$(1/A)_h$, l'inverse de la moyenne harmonique du nombre d'adultes par ménage

La probabilité de choisir un adulte de la strate h est dans ce cas $P_h=(1/M)\Sigma'1/A_{hi}$ $(i=1,...,N_h)$

Finalement, l'estimateur de la variance de y_{mp} suivant s'obtient donc à l'aide de la formule 2.3

$$\text{var}\,(y_{mp}) = \Sigma W_h^2((1/mP_h) + (1-P_h)/(m^2 P_h^2)) \text{ var } Y_h \qquad (h=1,...,N_h) \quad (2.4)$$

où $$\text{var } Y_h = \Sigma ((1/A)_h a_{hi} y_{hi}) - y_h)^2/(m_h-1) \qquad (i=1,...,m_h)$$

La quantité $(1/A)_h$ n'est habituellement pas connue et est un obstacle à l'emploi de y_{mp} pour estimer la moyenne. C'est pourquoi un nouvel estimateur sera développé dans la prochaine section.

3.0 LA PONDÉRATION

3.1 Définition

Dans le cadre de ce travail, la pondération sera utilisée pour écrire de façon unifiée différents types d'estimateurs. Cette technique consiste pour un plan d'échantillonnage quelconque à choisir des poids pour chacune des observations. Cette pondération sert par la suite à construire un estimateur dit pondéré, y_p, qui permet d'estimer la moyenne **Y**:

$y_p = \Sigma w_i y_i / \Sigma w_i$ où w_i est le poids de l'observation y_i, i=1,...,n

Dans le cas d'un échantillonnage stratifié, ces poids sont $w_{hi} = n N_h/(n_h N)$ pour tous les individus d'une strate h. Dans le cas d'un échantillonnage à probabilités variables où chaque unité i a une probabilité connue π_i d'être tirée, ces poids sont $1/\pi_i$. Nous reviendrons sur ces estimateurs dans les sections 3.3 et 3.4.

Les poids sont souvent aléatoires. Evidemment, les observations aussi sont aléatoires. Comme y_p est la somme du produit de ces variables aléatoires, il sera alors difficile d'estimer la variance de l'estimateur y_p . C'est pourquoi, le jackknife sera utilisé, comme nous le verrons, pour estimer la variance.

3.2 Le jackknife

3.2.1 Théorie classique

Nous simplifions ici la théorie du jackknife. Le lecteur pourra trouver une description plus détaillée de la technique générale dans l'article de Miller(1974). De la théorie générale du jackknife, g groupements d'observations de taille m peuvent être faits à partir d'un échantillon de taille n où n=mg. Dans ce texte, pour éviter le caractère arbitraire de la formation des groupes, nous considérons m=1. Le choix de g=n est des plus judicieux si l'on se fie à un article de Rao et Webster(1966). Ces derniers ont en effet montré à l'aide de simulations que la variance et le biais de l'estimateur du jackknife étaient une fonction décroissante de g. Définissons d'abord l'estimateur du jackknife.

Soit Θ, l'estimateur d'un paramètre θ basé sur un échantillon de taille n. Soit Θ_{-i}, l'estimateur de θ basé sur un échantillon de taille n-1 avec l'observation i en moins. Des pseudo-valeurs peuvent être définies par $\beta_i = n\,\Theta - (n-1)\,\Theta_{-i}$ $(i=1,\ldots,n)$.

La moyenne de ces pseudo-valeurs $\beta = (1/n)\sum \beta_i$ est appelée estimateur du jackknife et a la propriété d'éliminer le terme d'ordre 1/n du biais de la forme: $E[\Theta] - \theta = a_1/n + O(1/n^2)$

Selon un article de Tukey(1958), les pseudo-valeurs peuvent être supposées indépendantes et suivre une même loi dans plusieurs situations. Sous cette approximation, la variance de l'estimateur du jackknife peut être estimée par:

$$\text{var}\,\beta = \sum (\beta_i - \beta)^2 / n(n-1) \text{ ou encore}$$

$$= ((n-1)/n) \sum (\Theta_{-i} - \Theta_{(.)})^2 \text{ où } \Theta_{(.)} = \sum \Theta_{-i} / n$$

Sous cette hypothèse, selon cet article de Tukey, la statistique $(\beta - \theta) / (\text{var}(\beta))^{1/2}$ suit alors une t de Student avec n-1 degrés de liberté.

3.2.2 Application à l'estimateur pondéré

La précédente théorie du jackknife est maintenant appliquée à l'estimateur pondéré afin d'en estimer la variance qui est souvent difficile à calculer lorsque les poids sont aléatoires et de réduire le biais de l'estimateur pondéré s'il y en a. En effet, l'estimateur pondéré peut être considéré comme un estimateur quotient de la moyenne et il est ainsi souvent biaisé.

Pour simplifier la notation, on utilisera plutôt comme poids dans cette section seulement w_i / w où $w = \sum w_i /n$ et alors $y_p = \sum w_i y_i/n$.

Ici $\Theta = y_p = \sum w_i y_i / n$ et

$$\beta_i = n\,\Theta - (n-1)\,\Theta_{-i} = n\,y_p - (n-1) \sum w_j y_j / \sum w_j \quad (j \neq i)$$

et l'estimateur du jackknife qui sera noté y_j dans ce cas s'exprime comme

$$y_j = \sum \beta_i /n = n\,y_p - ((n-1)/n) \sum \sum w_j y_j / \sum w_j \qquad (i=1,\ldots,n)(j \neq i)$$

ou comme $y_j = \sum \beta_i /n = (1/n) \sum (n\,\Theta - (n-1)\,\Theta_{-i}) = n\,\Theta - (n-1)\,\Theta_{(.)}$

où $\Theta_{(.)} = \sum \Theta_{-i} / n = \sum (n y_p - w_i y_i)/(n-w_i) \qquad (i=1,2,\ldots,n)$

$$= (y_p \sum 1/(n-w_i)) - (1/n) \sum w_i y_i/(n-w_i)$$

L'estimateur de la variance de y_j s'obtient de la formule de var (β) et sera donc

$$\text{var}(y_j) = ((n-1)/n) \sum (\Theta_{-i} - \Theta_{(.)})^2 \qquad (i=1,2,...n)$$

$$= ((n-1)/n) \sum (((ny_p - w_i y_i)/(n-w_i)) - \Theta_{(.)})^2$$

Une formule simplifiée qui pourra être utilisée pour estimer la variance d'échantillonnage de y_p, notée var(y_p) peut être déduite de cette dernière quantité si l'on suppose que $\Theta_{(.)} \approx y_p$, c'est-à-dire que $y_p \approx \beta$.

Sous cette hypothèse, l'estimateur de la variance obtenu se déduit des approximations suivantes:

$$\text{var}(\beta) \approx ((n-1)/n) \sum (\Theta_{-i} - y_p)^2$$

$$= ((n-1)/n) \sum (((ny_p - w_i y_i)/(n-w_i)) - y_p)^2$$

(supposant $\sum w_i = n$)

$$= ((n-1)/n) \sum w_i^2 (y_i - y_p)^2 / (n-w_i)^2$$

$$\approx \sum w_i^2 (y_i - y_p)^2 / n(n-1) = \text{var}(y_p)$$

si l'on remplace $n-w_i$ par $n-1$.

Plus généralement, pour des w_i quelconque ($\sum w_i \neq n$), $\text{var}(y_p) = \sum (w_i/w)^2 (y_i - y_p)^2 / n(n-1)$. Cette dernière formule permet le calcul simplifié de la variance de y_p par le jackknife.

3.3 La pondération et la stratification

Soit le cas d'un tirage aléatoire simple dans L strates. Dans ce cas, la pondération est induite par le plan de sondage et permet d'estimer la moyenne par un estimateur sans biais. Dans ce cas particulier, les poids ne sont pas aléatoires comme nous le verrons.

Soit **Y**, une moyenne que l'on veut estimer. Dans le cas d'un plan stratifié à répartition non-proportionnelle, l'estimateur classique y_{st} est utilisé:

$$y_{st} = \sum W_h y_h \qquad (h=1,2,...,L)$$

où h ,la strate

y_h ,la moyenne de l'échantillon de la strate h

et $W_h = N_h / N$ où N_h, la taille de la population de la strate h

N , la taille de la population

La différence entre y et y_{st} est que dans le cas de l'estimateur y_{st} , les observations sont pondérées. Il est évident pour un même échantillon que y coincide avec y_{st} pourvu que $n_h/n = N_h/N$. L'obtention des poids de chacune des observations se fait tout simplement à l'aide d'une règle de trois. Il y a n_h observations dans la strate h et il en faudrait (nN_h/N) pour que l'échantillon soit parfaitement représentatif de la population. Il faudra alors $n_h\, w_h = n\, N_h/N$

Les poids associés à toutes les observations de la strate h, notés w_h, se calculent donc, lorsqu'on désire conserver la taille de l'échantillon original égal à n, par la formule

$$w_h = n\, N_h / n_h\, N$$

où n ,le nombre de répondants dans l'échantillon

n_h ,le nombre de répondants par strate dans l'échantillon.

Dans le cas de la stratification, il est très facile de démontrer l'égalité entre y_p et y_{st}. L'égalité des variances d'échantillonnage de y_p et y_{st} s'ensuit. Ainsi la variance de y_p pourra être estimée par l'estimateur sans biais

$$\text{var}(y_{st}) = \Sigma W_h^2 (\text{var } y_h)/n_h \quad (h=1,2,...,L)$$
$$\text{où var } y_h = \Sigma (y_{hi} - y_h)^2/(n_h-1) \quad (i=1,2,...,n_h)$$

Comme y_p est l'estimateur pondéré, il est intéressant de valider la théorie du jackknife décrite précédemment pour estimer la variance. Pour ce faire, nous allons donc montrer que l'approximation de l'estimateur de la variance du jackknife, var(y_p), qui est utilisée quand le plan est stratifié, est approximativement égale à l'estimateur de la variance de y_{st}.

Malheureusement, l'estimateur du jackknife ne prend pas en considération que le plan est stratifié lors du calcul de la variance. En effet, l'estimateur du jackknife utilisé pour estimer la variance ne considère pas que les moyennes des strates sont souvent différentes.

Afin de prendre en considération la stratification et pour obtenir un estimateur d'utilisation simple, nous allons soustraire de chaque observation la moyenne des strates associées et appliquer le jackknife sur les données réduites $y'_{hi} = y_{hi} - y_h$ tout en conservant les mêmes poids.

L'approximation de la variance de y_p déduite de la théorie du jackknife, s'écrit alors

$$\begin{aligned}
\text{var } y_p &= \Sigma \Sigma (w_h^2)(y'_{hi} - y'_p)^2/n(n-1) \quad (h=1,...,L)(i=1,...,n_h) \\
&= \Sigma \left((nN_h)/(Nn_h)\right)^2 \Sigma (y_{hi} - y_h)^2/n(n-1) \\
&\qquad \text{(puisque } y'_p=0) \\
&= \Sigma W_h^2 (\text{var } y_h/n_h)\, (n/n-1)((n_h-1)/n_h) \\
&\approx \Sigma W_h^2 \text{ var } y_h/ n_h = \text{var } y_{st}
\end{aligned}$$

3.4 La pondération et l'échantillonnage à probabilités variables

3.4.1 Un nouvel estimateur

Les estimateurs classiques utilisés suite à un tirage à probabilités variables ont été présentés dans la section 2.2.1. Dans cette section, l'estimateur de la moyenne y_{pv1} et l'estimateur de la variance de ce dernier var(y_{pv1}) ont été présentés. Le problème pratique de l'emploi de ces derniers estimateurs est qu'il nécessite la connaissance exacte de la taille de la population N et de la probabilité du tirage de chaque individu i parmi N. Dans le cadre de la pondération, on désire en quelque sorte remplacer les quantités π_i par des poids w_i.

Du chapitre 2, rappelons que l'estimateur de la moyenne dans le cas d'un échantillonnage à probabilités variables est

$$y_{pv1} = (1/nN) \Sigma y_i/\pi_i \quad (i=1,2,...,n)$$

Si l'on pose $w_i = 1/\pi_i$ dans l'équation de l'estimateur pondéré, on obtient un autre estimateur de la moyenne. Ce dernier est noté y_{pv2} quand le plan est à probabilités variables:

$$y_{pv2} = (\Sigma y_i/\pi_i) / \Sigma 1/\pi_i \quad (i=1,2,...,n)$$

Ce nouvel estimateur de la moyenne est différent de y_{pv1} en général. On aura l'égalité entre y_{pv1} et y_{pv2} si et seulement si $\Sigma\ 1/\pi_i = nN$ (comme par exemple, lorsque $\pi_i = 1/N$ pour toutes les observations. Ce nouvel estimateur a des propriétés différentes de y_{pv1}. D'abord y_{pv2} étant un estimateur quotient , il est biaisé. Il est facile de montrer que le premier terme de son biais est

$$\text{Biais}(y_{pv2}) = (-1/n)\ [(1/N^2)\ \Sigma\ (1/\pi_i)\ (y_i - Y)] \qquad (i=1,2,...,N)$$

De plus si l'on ajoute une constante c aux observations y_i, y_{pv2} sera augmenté de c tandis que y_{pv1} sera augmenté de $(c/nN)\ \Sigma(1/\pi_i)$. Il est donc très facile de montrer que la variance de y_{pv1} est sensible à une translation tandis que la variance de y_{pv2} ne l'est pas. En effet, l'estimation de la moyenne des observations translatées $y'_i = y_i + c$ s'exprime à l'aide des deux estimateurs comme

$$y'_{pv2} = \Sigma\ (y_i+c)/\pi_i\ /\ (\Sigma\ 1/\pi_i) = y_{pv2} + c$$

et

$$y'_{pv1} = (1/nN)\ \Sigma\ (y_i+c)/\pi_i\ = y_{pv1} + (c/nN)\Sigma\ 1/\pi_i$$

On en déduit aisément que $\text{Var}\ y'_{pv2} = \text{Var}\ y_{pv2}$ tandis que $\text{Var}\ y'_{pv1} \neq \text{Var}\ y_{pv1}$. D'un article de Dayal (1981,p166), la variance de y_{pv2} peut être déduite. Dans cet article, la variance de y_{pv2} est égale en ignorant le facteur de population finie à

$$\text{Var}\ (y_{pv2}) = (1/nN)\ \Sigma'(W_i/W)^2(Y_i - Y_p)^2 \qquad (i=1,2,...,N)$$

où $W_i = 1/\pi_i$, $Y_p = \Sigma' W_i Y_i / \Sigma\ W_i$ et $W = (1/N)\ \Sigma'(1/\pi_i)$

La variance de y_{pv1} peut être beaucoup plus grande que la variance de y_{pv2} comme en fait foi une des simulations fournies dans le tableau 1 ayant pour but d'estimer la moyenne d'une normale centrée en 50 et de variance 1. Chaque simulation est le produit de 1000 tirages avec remise de 100 personnes dans une population de 1000 personnes. Il y a dans cette population 5 groupes de 200 personnes dont les probabilités de sélection de chaque membre d'un groupe à un tirage sont respectivement de 0,0005, 0,001, 0,00075, 0,00125 et 0,0015. Ces simulations ont pour objet d'estimer les paramètres p pour deux lois de Bernoulli B(0,05) et B(0,5), et à estimer la moyenne de trois normales N(0, 1), N(50, 1) et N(50, 50).

Des simulations présentées dans le tableau 1, deux résultats retiennent notre attention. D'abord le biais théorique est très faible et peut être considéré négligeable. Le biais théorique ne peut être détecté à l'aide de ces simulations puisque la variabilité échantillonnale est beaucoup plus grande que le biais théorique. Dans le cas d'un échantillonnage aléatoire simple de 1000 échantillons de 100 personnes, la marge d'erreur d'échantillonnage serait de 0,003.

En second lieu, on remarque la variabilité qu'on peut qualifier d'explosive de y_{pv1}. Lors de l'estimation de la moyenne d'une N(50, 1), la variabilité de y_{pv1} est 355 fois supérieure à celle de y_{pv2} et dans le cadre de l'estimation d'une N(50, 50), elle est près de 9 fois supérieure. C'est pourquoi nous favoriserons y_{pv2} lorsque le tirage sera fait à probabilités variables.

De l'équation précédente, qui fournit une expression de la variance de y_{pv2}, il est très facile de montrer que la variance d'échantillonnage de y_{pv2} peut être estimé par:

$$\text{var}\ (y_{pv2}) = (1/n(n-1))\ \Sigma\ (w_i/w)^2(y_i - y_{pv2})^2 \text{ où } w_i = 1/\pi_i.$$

Ce dernier estimateur de la variance est identique à var(y_p). Ceci confirme la qualité de l'estimateur du jackknife pour estimer la variabilité de l'estimateur pondéré dans le cas d'un tirage à probabilités variables.

Tableau 1: Estimation de la moyenne et de la variance de y_{pv1} et y_{pv2} suite à la simulation de 1000 tirages à probabilités variables de 100 personnes.

	B(0,5)	B(0,05)	N(0,1)	N(50,1)	N(50,50)
Moyenne					
y_{pv1}	0,4980	0,05010	0,00157	50,044	49,967
y_{pv2}	0,4975	0,05006	0,00185	50,003	49,981
Variance					
y_{pv1}	0,00296	0,0006508	0,01204	4,1637	4,6747
y_{pv2}	0,00261	0,0006406	0,01196	0,0117	0,5437
Biais*	$-1,33\times10^{-6}$	$-3,06\times10^{-5}$	$1,67\times10^{-4}$	$1,67\times10^{-4}$	$1,18\times10^{-3}$

* Ce biais est calculé à l'aide de la formule pour Biais(y_{pv2})

De l'équation précédente, qui fournit une expression de la variance de y_{pv2}, il est très facile de montrer que la variance d'échantillonnage de y_{pv2} peut être estimé par:

$$\text{var}(y_{pv2}) = (1/n(n-1)) \sum (w_i/w)^2 (y_i - y_{pv2})^2 \text{ où } w_i = 1/\pi_i.$$

Ce dernier estimateur de la variance est identique à var(y_p). Ceci confirme la qualité de l'estimateur du jackknife pour estimer la variabilité de l'estimateur pondéré dans le cas d'un tirage à probabilités variables.

3.4.2 Application au tirage d'un seul adulte par ménage

A titre d'exemple, la théorie du jackknife est appliquée à l'échantillonnage d'adultes dans des ménages qui sont retenus suite à un échantillonnage aléatoire simple. Dans les ménages retenus, il y a sélection d'un seul adulte par ménage. Ce cas, discuté précédemment, est un cas d'échantillonnage à probabilités variables. La notation présentée dans la section 2.2.1 est celle qui sera utilisée ici.

Comme dans l'exemple du plan stratifié, la pondération revient à identifier n poids w_i qui cherchent à rendre l'estimateur y_p non-biaisé. Dans le cas simple de la sélection d'un adulte par ménage où chaque ménage a même probabilité d'être sélectionnée, la probabilité de sélection est directement proportionnelle à l'inverse du nombre d'adultes dans le ménage. Les poids seraient dans un tel cas a_i/a puisque $\pi_i = 1/a_i M$ où a_i est le nombre d'adultes du i ème ménage et M, le nombre de ménages dans la population..

En posant w_i égal à a_i/a dans y_p, on obtient ainsi l'estimateur pondéré:

$$y_{m2} = (1/am) \sum a_i y_i \quad (i=1,2,...,m)$$

où **a** est le nombre moyen d'adultes par ménage de notre échantillon.

On vérifie que ce dernier estimateur est presque identique à y_{m1}. La seule différence est que dans cette dernière formule A est remplacé par **a**. Or , **a** est un estimateur non-biaisé de **A**. L'estimateur y_{m2} semble donc tout à fait indiqué pour estimer **Y**. Celui-ci est toutefois biaisé.

Mais asymptotiquement, ce biais est négligeable. Il est facile, en décomposant y_{m2} sous la forme de série de Taylor, de montrer que y_m est biaisé d'ordre 1/m.

Pour estimer la variance de l'estimateur y_{m2}, on a vu en 3.4.1 que l'estimateur var y_p issu de la théorie du jackknife est l'estimateur suggéré. Dans ce cas, il s'exprime comme

$$\text{var } y_{m2} = \Sigma [(a_i/a)^2 (y_i - y_{m2})]^2 /(m(m-1))$$

4.0 SIMULATIONS EN VUE D'ÉVALUER UN INTERVALLE DE CONFIANCE

Ce chapitre porte sur des simulations de Monte Carlo programmées en Pascal sur Macintosh II. Ces simulations ont pour but de valider, dans le cas de deux méthodes d'échantillonnage, la couverture de certains paramètres des lois les plus usuelles. Les lois simulées sont Bernoulli, normale, log-normale et khi-carré. Dans les simulations, le risque d'erreur α est fixé à 5%, le risque d'erreur usuel. L'intervalle de confiance basé sur la formule simplifiée d'estimation de la variance, $[y_p \pm t_{(n-1,1-\alpha/2)} (\text{var}(y_p))^{0,5}]$, est mis à l'essai de simulations de Monte-Carlo en vue d'évaluer si le seuil de confiance observé est près de 95%. L'interprétation de cet intervalle est que si l'on tire un grand nombre d'échantillons et que l'on calcule cet intervalle pour chaque échantillon, alors approximativement $100(1-\alpha)\%$ de ces intervalles contiendront la valeur du paramètre à estimer dans la population. Bien sûr, en posant cet intervalle de confiance, on se base évidemment sur le théorème limite central et nous allons vérifier en pratique dans ce chapitre qu'il s'applique dans le cas de l'échantillonnage stratifié à répartition non-proportionnelle et dans le cas d'un échantillonnage simple avec post-stratification. Il va de soi que l'aspect de la population joue un rôle très important. Si l'asymétrie de la distribution est très prononcée dans une population, la probabilité de couverture aura tendance à s'écarter de la valeur théorique, $(1-\alpha)$.

4.1 Cas d'un tirage stratifié à répartition non-proportionnelle

Pour valider la précédente théorie, nous allons considérer le tirage de 100 personnes dans chacune des 6 strates d'une population de 12000 personnes. La fiabilité des trois intervalles sera testée au moyen de simulations par la méthode de Monte Carlo, c'est-à-dire en tirant mille échantillons, chacun dans ce cas-ci selon le plan stratifié prédéfini, et en calculant la proportion des 1000 intervalles qui inclut la vraie valeur du paramètre à mesurer. Cette proportion devrait s'approcher de 95%.

Dans la population, 6 paramètres sont à estimer. Deux de ces paramètres sont les paramètres p de deux lois de Bernoulli de paramètre p_1=0,5 et p_2=0,04. La distribution de ces proportions à l'intérieur de chacune des strates est présentée dans le tableau 2. Les troisième et quatrième paramètres à estimer sont les moyennes μ_1=0 et μ_2=50 de deux normales $N(\mu_1,1)$, $N(\mu_2,1)$. Le cinquième paramètre à estimer est la moyenne d'une khi-carré d'un degré de liberté, i.e. 1. Finalement, le dernier paramètre à estimer sera la moyenne d'une log-normale dont le premier moment est $e^{.5}(1.6472)$. La distribution log-normale est obtenue en prenant l'exponentielle à base e de la distribution normale centrée réduite et la khi-carré est obtenue en prenant le carré des observations de la distribution normale. La distribution normale centrée réduite est générée par la méthode de Box-Muller. Ces deux dernières distributions et la Bernoulli de paramètre p_2=0,04 ont été choisies pour leurs asymétries connues.

Tableau 2. Distributions par strate des paramètres p_1 et p_2 de deux lois de Bernoulli.

Strate h	N_h	W_h	$B(p_1)$	$B(p_2)$
1	1000	1/12	50%	5%
2	1000	1/12	50%	3%
3	2000	1/6	30%	6%
4	2000	1/6	70%	2%
5	3000	1/4	60%	4%
6	3000	1/4	40%	4%

Dans ces simulations et les prochaines, il y a indépendance entre les simulations, i.e. qu'il y a répétition pour chaque loi du tirage de 1000 échantillons.

A l'intervalle présenté, on ajoute dans les simulations, pour le comparer, l'intervalle dit classique $[y_{st} \pm t_{n-1,(1-\alpha)/2}(\text{var } y_{st})^{0,5}]$. Il a déjà été montré que y_{st} est égal à l'estimateur pondéré y_p et que l'estimateur de la variance $\text{var}(y_p)$ est presque égal à $\text{var } y_{st}$. C'est pourquoi on n'attend pas, de cet intervalle, un seuil de confiance différent des deuxième et troisième intervalles.

Les résultats de ces simulations sont fournis au tableau 3. Dans ces simulations, il n'y a pas de différence importante entre les deux estimateurs de variances retenus. Le taux de rejet des deux intervalles retenus est ainsi presque identique pour toutes les distributions. Dans une simulation, le taux de rejet est la proportion des intervalles qui ne contiennent pas la valeur du paramètre à estimer. Comme prévu, le taux de rejet de l'intervalle classique est très près du taux de rejet de l'intervalle $y_p \pm 1{,}96\ (\text{var}(y_p))^{1/2}$.

En théorie, en considérant que le taux de rejet bilatéral théorique est de 5% et qu'il y a 1000 tirages, la proportion des rejets bilatéraux devrait se situer entre 0,0365 et 0,0635 dans 95% des cas. Ce dernier intervalle est obtenu en considérant qu'une loi binomiale de paramètre n=1000 et p=0,05 peut être approchée par une loi normale de moyenne np et de variance npq. Un seul écart significatif surprenant entre le taux de rejet bilatéral et le taux théorique, 5%, avec un seuil de confiance de 95% est observé lorsque la loi est Bernoulli (0,5). Ceci est surprenant puisque cette distribution peut être qualifiée de symétrique.

4.2 Cas d'une post-stratification suite à un échantillonnage simple

Lors des simulations précédentes, le plan d'échantillonnage était stratifié. Le nombre d'unités à choisir dans chaque strate était fixé à l'avance à 100 unités. Lors d'une post-stratification, on fixe plutôt le nombre total d'unités à obtenir, dans notre cas à 600, et à la fin les unités obtenues par un tirage aléatoire simple sont classées dans des strates que l'on appelle post-strates. La population dans laquelle les échantillons sont tirés est identique à celle présentée dans le tableau 2.

La fiabilité des intervalles est testée encore cette fois-ci au moyen de simulations par la méthode de Monte Carlo. Les paramètres des distributions présentées dans la section précédente sont encore ceux à estimer. A l'intervalle déja choisi, on ajoute, pour le comparer, l'intervalle classique $[y_{pst} \pm t_{n-1,1-\alpha/2}\ (\text{var } y_{pst})^{0,5}]$.

Les résultats des simulations sont fournis dans le tableau 4. Dans ce tableau, les taux de rejet des deux intervalles de confiance sont presques identiques pour tous les intervalles. Il n'y a aucun écart significatif entre la proportion estimée de rejet bilatéral et le taux théorique de 5%.

Tableau 3. Résultats des simulations par la méthode de Monte Carlo de tirages de cent personnes dans 6 strates.

	B(0.5)	B(0.04)	N(0,1)	N(50,1)	LN	$X^2(1)$
Estimateur de la moyenne						
- moy (y_{st})	,50098250	,03984083	-,0025179	49,9989689	1,64563	1,001854
- moy (y_p)	,50098250	,03984083	-,0025179	49,9989689	1,64563	1,001854
Estimateur de la variabilité des estimateurs						
- moy (var (y_{st}))	,00045107	,00007439	,00195060	,00193627	,008761	,0039701
- moy (var(y_p))	,00044731	,00007377	,00193432	,00192011	,008688	,0039370
Intervalle classique						
-% rejet	6,7	6,1	5,1	5,9	6,1	5,5
Intervalle $y_p \pm 1{,}96\ (\text{var}(y_p))^{1/2}$						
-% rejet	7,0	6,2	5,1	5,9	6,1	5,5

Tableau 4 Résultats de simulations par la méthode de Monte Carlo de post-stratification selon 6 strates suite à un tirage aléatoire simple de 600 personnes.

	B(0.5)	B(0.04)	N(0,1)	N(50,1)	LN	$X^2(1)$
Estimateur de la moyenne						
- moy (y_{pst})	,4996899	,04007913	,00284426	50,0008310	1,65058	1,000252
- moy (y_p)	,4996899	,04007913	,00284426	50,0008310	1,65058	1,000252
Estimateur de la variabilité des estimateurs						
- moy (var (y_{pst}))	,0003925	,00006430	,00167895	,00168064	,007427	,0033828
- moy (var(y_p))	,0003868	,00006387	,00166980	,00166722	,007416	,0033379
Intervalle classique						
-% rejet	4,8	4,2	5,5	4,5	4,8	5,8
Intervalle $y_p \pm 1{,}96\ (\text{var}(y_p))^{1/2}$						
-% rejet	5,0	4,2	5,7	4,5	4,2	5,8

En conclusion, les estimateurs de la moyenne et de la variance issus de la théorie du jackknife diffèrent généralement très peu de ceux qui proviennent de la théorie classique quand l'échantillonnage est stratifié et quand il y a post-stratification.

5.0 ESTIMATION DE LA MOYENNE ET DE LA VARIANCE LORS DE L'ENQUÊTE SOM-R

Dans le cadre du Som-R, les échantillons recueillis ne sont pas représentatifs de la population sondée. C'est pourquoi, il faut envisager la post-stratification ou la pondération. Usuellement, une post-stratification est produite par rapport à l'âge et au sexe dans chacune des trois régions de cueillette. Pour estimer la moyenne et la variance d'échantillonnage d'observations issues de la cueillette du Som-R en considérant la post-stratification, on utilisera deux moyens dérivés de ce qui a été vu précedemment. Le premier moyen d'estimer ces quantités consiste à développer la théorie classique vue à la section 2.2.2 sur la post-stratification suite à une méthode d'échantillonnage à probabilités variables au cas du Som-R. Le second moyen sera d'identifier la bonne pondération et d'estimer la moyenne par l'estimateur pondéré et la variance d'échantillonnage de l'estimateur pondéré par $var(y_p)$.

5.1 Par post-stratification

Le nombre d'entrevues à compléter dans chaque région est connu à l'avance et est noté m_r, le nombre de ménages à choisir dans la région r. Suite aux tirages à probabilités variables d'adultes dans une région, la post-stratification est envisagée. La théorie présentée à la fin de la section 2.2.2 s'applique au traitement d'une seule région. La lettre r est ajoutée à la notation présentée dans cette section pour indiquer l'appartenance d'un adulte à une région. L'estimateur de la moyenne des observations de la région r est déduit de cette section et sera noté y_r. La valeur d'une observation du i ème adulte de la population de la post-strate h qui réside dans la région r est notée Y_{rhi}, $i=1,2,...,N_{rh}$ et le nombre d'adultes présents dans le ménage de cet adulte est A_{rhi}.

L'estimateur post-stratifié de la moyenne des observations d'une région est donc

$$y_r = \sum W_{rh}\, y_{rh} \qquad (h=1,2,...,L)$$

où

$$y_{rh} = (1/A)_{rh} \sum a_{rhi}\, y_{rhi} / m_{rh} \qquad (i=1,2,...,m_{rh})$$

où

m_{rh}, le nombre des ménages parmi les m_r retenus dans la région r dont l'adulte choisi est classé dans la strate h

y_{rhi}, la valeur d'une observation de l'échantillon parmi m_{rh}

a_{rhi}, le nombre d'adultes dans le i ème ménage retenu de la région r dont l'adulte choisi est classé dans la strate h

$(1/A)_{rh} = \sum(1/A_{rhi})/N_{rh}$

W_{rh}, la proportion d'adultes de la région r qui sont de la strate h $(W_{rh}=N_{rh}/N_r)$

Dans la région r, la probabilité de choisir un adulte de la strate h est

$$P_{rh} = (1/M_r) \sum 1/A_{rhi} \quad (i=1,2,...,N_{rh})$$

où M_r est le nombre de ménages dans la région r

L'estimateur de la variance d'échantillonnage de y_r est alors d'après la formule 2.3

$$var(y_r) = \sum W_{rh}^2((1/m_r P_{rh}) + (1-P_{rh})/(m_r^2 P_{rh}^2))\, var Y_{rh} \qquad (5.1)$$

où

$$var\, Y_{rh} = \sum ((1/A)_{rh}\, a_{rhi}\, y_{rhi}) - y_{rh})^2 / (m_{rh}-1) \qquad (i=1,...,m_{rh})$$

Les quantités $(1/A)_{rh}$ et P_{rh} sont difficilement connues par la voie des recensements et sont un obstacle majeur à l'emploi de ces derniers estimateurs. Par exemple, s'il y a une post-stratification en fonction du sexe et des groupes d'âge 18 à 24 ans, 25 à 34 ans, etc, il serait très difficile de connaître la moyenne de l'inverse du nombre d'adultes par ménage de la région r qui contient un ou des hommes âgés entre 25 et 34 ans.

Finalement, l'estimateur de la moyenne Y d'observations de la population du Québec sera, lorsque l'on connaît exactement les quantités $(1/A)_{rh}$ et P_{rh}, l'estimateur stratifié $y_{S1} = \Sigma W_r y_r$ (r=1,2,3), où W_1 est la proportion d'adultes de la province de Québec qui résident dans la région métropolitaine de Québec, W_2 est la proportion de ceux qui résident dans la région métropolitaine de Montréal et W_3 est la proportion de ceux qui résident ailleurs que dans ces deux régions métropolitaines. Les échantillons sont obtenus de tirages indépendants dans les différentes régions. Ainsi l'estimateur de la variance de y_{S1} sera $\text{var}(y_{S1}) = \Sigma W_r^2 \text{var}(y_r)$.

5.2 Par pondération

La pondération consiste à la recherche de poids qui vont tenter de rendre l'estimateur pondéré non-biaisé. La recherche de ces poids peut paraître difficile à première vue. La solution est simplement d'associer à chaque source de non-représentativité une pondération. La pondération s'exprime par la suite comme le produit des pondérations de chaque source.

Pour le Som-R, à première vue, trois pondérations sont nécessaires. Une première pondération permet de considérer que le plan est à probabilités variables. Une deuxième est utile à la post-stratification et une dernière est utilisée pour considérer la stratification par région. La stratification est simple à traiter par pondération, mais il faut considérer lors de l'estimation de la variance d'échantillonnage de l'estimateur pondéré que les moyennes des strates sont par hypothèse différentes. La post-stratification peut être considérée comme une stratification et il faudra donc aussi centrer les données dans chacune des strates. La pondération finale sera le produit de ces trois pondérations. Puisque la stratification a asymptotiquement le même effet que la post-stratification, il est sensé de considérer ces deux pondérations simultanément ou comme une seule pondération par région et par post-strate.

La première pondération se doit de corriger l'effet du tirage d'un seul répondant par ménage. Une telle pondération a déjà été vue à la section 3.4.2. On a déjà vu que le poids d'un individu se doit d'être inversement proportionnel à sa probabilité de sélection qui elle même est déjà inversement proportionnelle au nombre d'adultes dans le ménage, a_{rhi}. Définissons d'abord, a_{rh}, le nombre moyen d'adultes par ménage dont l'adulte de l'échantillon est de la strate h et de la région r. Par référence à la section 3.4.2, l'estimation de la moyenne à l'intérieur d'une post-strate et d'une région r s'exprime par

$$y_{rh} = \Sigma a_{rhi} y_{rhi} / (a_{rh} m_{rh}) \qquad (i=1,2,..,m_{rh})$$

Ainsi la première pondération sera a_{rhi}/a_{rh}. La quantité y_{rh} sera utile pour centrer les données des post-strates avant d'estimer la variance puisqu'il y a stratification et post-stratification.

Lors d'une post-stratification et d'une stratification , le calcul des poids se fait comme dans le cas d'une stratification. Après tirage, il y a m_{rh} observations dans la post-strate h de la région r et il en faudrait mN_{rh}/N pour que l'échantillon soit représentatif de la population. Le poids d'un adulte de la post-strate h et de la région r doit donc être proportionnel à N_{rh}/m_{rh} d'après la section 2.4.

Le poids final d'une observation classée dans une post-strate h et provenant d'une région r s'exprime finalement par le produit des poids N_{rh}/m_{rh} et a_{rhi}/a_{rh}.

L'estimateur pondéré de la moyenne associé à l'enquête du Som-R sera donc finalement

$$y_{S2} = \Sigma\Sigma\Sigma w_{rhi}\, y_{rhi} / \Sigma\Sigma\Sigma w_{rhi} \qquad (r=1,2,3),(h=1,...,L),(i=1,...,m_{rh})$$
$$\text{où } w_{rhi} = N_{rh}\, a_{rhi} / a_{rh}\, m_{rh}$$

Puisqu'il y a stratification et post-stratification, il faut estimer la variance de l'estimateur du jackknife à partir d'observations réduites par rapport aux régions et aux post-strates.

L'estimateur de la variance de l'estimateur pondéré du Som-R s'exprimera donc comme

$$\text{var}(y_{S2}) = \Sigma\Sigma\Sigma (w_{rhi}/w)^2 (y_{rhi} - y_{rh})^2/(m(m-1)) \qquad (r=1,2,3),(h=1,...,L),(i=1,...,m_{rh})$$

$$\text{où } w = \Sigma\Sigma\Sigma\, w_{rhi}/m \qquad (r=1,2,3),\ (h=1,...,L),\ (i=1,...,m_{rh})$$

5.3 Comparaison des deux intervalles de confiance

Les simulations présentées dans cette section vont permettre de comparer les deux estimateurs y_{S1} et y_{S2}. La couverture des deux intervalles de confiance créés à partir de ces estimateurs sera testée par la méthode de Monte Carlo. C'est-à-dire en tirant mille échantillons selon la méthode d'échantillonnage du Som-R, et en calculant la proportion des 1000 intervalles qui incluent la vraie valeur du paramètre à mesurer. Cette proportion devrait s'approcher de 95%.

La population qui fera l'objet du tirage est représentée par 25000 adultes de la province de Québec qui sont répartis proportionnellement à partir des distributions d'âge et de sexe des régions métropolitaines de Montréal, de Québec et du reste de la province. Ces distributions proviennent du recensement canadien de 1986 et seront utiles à la post-stratification. Celles-ci sont présentées dans le tableau 6.

Pour calculer y_{S1}, les quantités $(1/A)_{rh}$ doivent être connues. Ces dernières ne le sont malheureusement pas. C'est pourquoi, des nombres moyens d'adultes par ménage, A_{rh}, sont choisis entre 1,5 et 2,5. Par la suite, un algorithme simple est utilisé pour distribuer les tailles des ménages à partir de ces tailles moyennes. De ces tailles, les quantités $(1/A)_{rh}$ sont calculées. Les statistiques A_{rh} et $(1/A)_{rh}$ sont fournies dans le tableau 6.

Dans le but d'imiter le plan d'échantillonnage du Som-R, le tirage effectué consiste à tirer 300 adultes dans le Québec métropolitain, 400 dans le Montréal métropolitain et 400 dans le reste de la province. La post-stratification par rapport à 6 groupes d'âge et au sexe est par la suite envisagée à partir des distributions d'âge et de sexe dans ces régions. Dans une région r, chaque adulte est classé dans un des 6 groupes d'âge et selon son sexe. Il y a donc en tout 12 post-strates par région.

Dans la population de 25000 personnes, 6 paramètres sont à estimer. Deux de ces paramètres sont les paramètres p1 et p2 de deux lois de Bernoulli 0,5 et 0,045676. La distribution de ces paramètres à l'intérieur des régions et selon les groupes d'âge et le sexe est fournie dans le tableau 7. Le troisième paramètre à estimer est la moyenne μ=0 d'une normale $N(\mu,1)$. Un quatrième paramètre est obtenu après l'ajout à la normale N(0,1) de constantes qui varient selon les post-strates et les régions. Ces constantes sont présentées dans le tableau 7. Les deux derniers paramètres à estimer sont aussi déduits de la normale centrée réduite. Une distribution log-normale est obtenue en prenant l'exponentielle à base e de la normale centrée

réduite. La moyenne de la distribution log-normale est le cinquième paramètre à estimer. Une distribution khi-carré est obtenue en prenant le carré de la distribution normale et son premier moment , i.e. 1, est le sixième et dernier paramètre que l'on désire estimer.

Tableau 6. Quelques statistiques sur la population utiles aux simulations de l'enquête Som-R

N_{rh}	Québec métro.		Montréal métro.		Ailleurs en province	
N_{rh}	Homme	Femme	Homme	Femme	Homme	Femme
A_{rh}						
$(1/A)_{rh}$						
18-24	186	183	902	896	911	875
	0,74%	0,73%	3,61%	3,58%	3,64%	3,50%
	2,0	2,3	1,8	1,9	2,2	2,0
	0,576	0,526	0,681	0,641	0,549	0,595
25-34	286	298	1363	1399	1382	1376
	1,14%	1,19%	5,45%	5,59%	5,53%	5,55%
	1,8	1,9	2,2	2,0	2,4	2,8
	0,683	0,642	0,550	0,595	0,505	0,415
35-44	245	256	1108	1162	1382	1376
	0,98%	1,02%	4,43%	4,65%	4,55%	4,44%
	2,2	2,0	2,4	2,8	2,15	1,95
	0,549	0,594	0,505	0,414	0,561	0,617
45-54	155	164	813	851	754	758
	0,62%	0,65%	3,25%	3,41%	3,02%	3,03%
	2,4	2,8	2,15	1,95	2,0	2,3
	0,505	0,418	0,561	0,618	0,595	0,527
55-64	123	147	688	779	667	722
	0,49%	0,59%	2,75%	3,12%	2,67%	2,89%
	2,15	1,95	2,0	2,3	1,8	1,9
	0,566	0,617	0,595	0,527	0,685	0,640
65 et +	107	182	584	938	653	849
	0,43%	0,73%	2,34%	3,75%	2,61%	3,40%
	2,0	2,3	1,8	1,9	2,2	2,0
	0,597	0,526	0,686	0,640	0,550	0,595

Le premier intervalle de confiance est extrait de la théorie classique qui considère que la post-stratification est appliquée suite à un tirage à probabilités variables. Cette théorie a été énoncée au début de ce chapitre. Le premier intervalle sera donc dans ce premier cas $[y_{S1} \pm 1.96(\text{var}(y_{S1}))^{.5}]$.

Le deuxième intervalle considéré dans ces simulations est déduit de la technique de pondération. L'estimateur pondéré obtenu a été présenté dans la deuxième section de ce chapitre. Rappelons que suite à la théorie présentée dans ce cas, l'estimateur de la moyenne d'observations provenant de l'enquête Som-R est y_{S2} et que l'estimateur de sa variance est noté $\text{var}(y_{S2})$. Le deuxième intervalle sera donc $[y_{S2} \pm 1.96(\text{var}(y_{S2}))^{.5}]$

Tableau 7. Distributions des paramètres p_1 et p_2 de deux lois de Bernoullis et des constantes k ajoutées à la loi normale centrée réduite par groupe d'âge et par sexe utiles aux simulations de l'enquête Som-R.

	Québec métro		Montréal métro		Ailleurs en province	
p_1						
p_2	Homme	Femme	Homme	Femme	Homme	Femme
k						
18-24	0,4	0,6	0,55	0,45	0,45	0,55
	0,03	0,05	0,05	0,03	0,03	0,03
	35	40	25	30	40	40
25-34	0,5	0,5	0,5	0,5	0,4	0,6
	0,07	0,02	0,04	0,05	0,03	0,03
	45	50	35	40	35	35
35-44	0,3	0,7	0,3	0,7	0,5	0,5
	0,12	0,1	0,03	0,06	0,03	0,03
	50	55	40	45	35	35
45-54	0,7	0,3	0,7	0,3	0,3	0,7
	0,03	0,06	0,03	0,05	0,06	0,06
	55	60	45	50	40	40
55-64	0,5	0,5	0,5	0,5	0,7	0,3
	0,05	0,04	0,04	0,04	0,07	0,08
	60	65	50	55	40	50
65 et +	0,45	0,55	0,6	0,4	0,5	0,5
	0,03	0,06	0,05	0,05	0,07	0,07
	65	70	55	60	50	50

Les résultats des simulations par la méthode de Monte Carlo sont présentés dans le tableau 8. Dans ces simulations, la moyenne des 1000 estimateurs y_{S1} et y_{S2} sont très près des valeurs réelles des paramètres dans la population. Mais, en moyenne l'estimateur y_{S1} est plus près que y_{S2} de la valeur réelle du paramètre dans 5 cas sur 6. Cela n'est sûrement pas attribuable au hasard et est dû au fait que y_{S2} est légèrement biaisé. Toutefois, l'estimateur y_{S2} varie moins dans 5 des 6 simulations et il varie beaucoup moins lorsque la distribution est normale centrée en 42,3795. Dans ce cas, la moyenne des mille estimateurs de la variance de y_{S1}, var y_{S1}, est plus de 300 fois plus élevée que la moyenne des mille estimateurs de variance de y_{S2}. Un cas similaire a déjà été vu dans la section 3.4.1 et n'est pas surprenant.

Les taux de rejet bilatéral des deux intervalles sont presque identiques. Des bonnes performances, i.e. des taux de rejet observés près des taux de rejet théoriques, sont présents quand les distributions sont symétriques. Les taux de rejet bilatéral sont tous supérieurs à 5%. En théorie pour le rejet bilatéral, en considérant que le taux de rejet bilatéral est de 5% et qu'il y a 1000 répétitions du Som-R, les taux observés de rejet bilatéral devraient se situer entre 0,0365 et 0,0635 dans 95% des cas. Des écarts significatifs avec le taux de rejet bilatéral de rejet et 5% sont observés quand les distributions sont Bernoulli de paramètre (0,04567), log-normale et khi-carré.

Tableau 8. Résultats de simulations par la méthode de Monte-Carlo de tirages par la méthode d'échantillonnage de l'enquête Som-R

	B (,50008)	B (,045676)	Norm (0,1)	Norm (42,484)	LN (1,655865)	X^2 (1,000)
Estimateur de la moyenne						
- moy (y_{S1})	,50044	0,04575	0,000029	42,477	1,655871	0,99988
- moy (y_{S2})	,50088	0,04569	-,000168	42,485	1,655846	0,99959
Estimateur de la variabilité						
- moy (var(y_{S1}))	,0003535	,0000561	,00129	,39053	,00707	,00281
- moy (var(y_{S2}))	,0002918	,0000540	,00135	,00126	,00629	,00252
Intervalle du jackknife						
-% rejet	6,2	7,4	5,6	4,4	6,4	6,5
Intervalle théorique						
-% rejet	6,3	6,8	5,5	5,8	6,6	6,5

CONCLUSION

La technique de pondération aurait avantage à être utilisée. Rapide à utiliser et facilement adaptable à plusieurs méthodes d'échantillonnage, la pondération produit des résultats très intéressants. Soit y_i , n observations pour i=1,2,..n et w_i les poids associés à ces observations qui dépendent de la méthode d'échantillonnage utilisée. L'estimateur $y_p=\Sigma w_i y_i/\Sigma w_i$ est asymptotiquement non-biaisé sa variance peut être approchée par la quantité $var(y_p)= \Sigma(w_i/w)^2(y_i-y_p)^2/(n(n-1))$ quand le plan est à probabilités variables. Si toutefois l'échantillonnage est stratifié ou s'il y a post-stratification, il faudra soustraire des observations l'estimation de la moyenne des strates ou des post-strates avant d'utiliser cette dernière formule. Lorsque l'échantillonnage est à probabilités variables, l'estimateur pondéré de la moyenne y_p est biaisé contrairement à l'estimateur classique y_{pv1} mais ce biais est très faible. Dans le cas d'un échantillonnage à probabilités variables, l'estimateur pondéré y_p varie beaucoup moins pour certaines distributions que l'estimateur classique non-biaisé y_{pv1}. Lorsque l'échantillonnage est stratifié ou s'il y a post-stratification, des simulations par la méthode de Monte Carlo, où 600 personnes étaient choisies, ont permis de montrer que le seuil de confiance de l'intervalle $y_p \pm 1.96\ (var(y_p))^{0,5}$ se situait près de 95% pour deux lois de Bernoulli, deux lois normales, une loi log-normale et une khi-carré.

L'application de la théorie classique de l'estimation dans l'enquête Som-R a conduit au développement d'un estimateur non-biaisé de la moyenne dont l'emploi nécessite la connaissance de quantités qui ne sont pas disponibles facilement. D'autre part, la technique de pondération conduit à un estimateur de la moyenne très peu biaisé et qui ne nécessite que de l'information facilement disponible. Très facile à utiliser dans l'enquête Som-R, l'estimateur pondéré est beaucoup moins variable que l'estimateur non-biaisé pour certaines distributions. Pour l'enquête du Som-R, des simulations par la méthode de Monte Carlo ont montré que le seuil de confiance de l'intervalle $y_p \pm 1.96(var(y_p))^{0,5}$ s'approchait de 95%.

BIBLIOGRAPHIE

BETHLEHEM, J.G. (1984). Linweight: A programm For Weighting Sample Survey Data. Compstat, 197-202.

* COCHRAN, W. G. (1977). Sampling Techniques. (3 rd ed.). New-York: John Wiley and Sons Inc.

CONVERSE, P.E., Traugott, M.W. (1986). Assessing The Accuracy Of Polls And Surveys. Science, 234, 1094-1098

* DAYAL, S. (1981). Sampling For Estimating Weighted Totals And Averages. Ann. Inst. Statist. Math, 33A, 165-176.

DOSS, D.C., HARTLEY, H.O. & SOMAYAJULU, G.R. (1979). An Exact Small Sample Theory For Post-Stratification. Journal of Statistical Planning And Inference, 3, 235-247.

EFRON,B., STEIN, C. (1981). The Jackknife Estimate Of Variance. Annals of Statistics, 9,586-596.

FULLER, W. A. (1966). Estimation Employing Post-Strata. American Statistical Association Journal, December, 1172-1183.

HINKLEY,D.V. (1977). Jackknife Confidence Limits Using Student t Approximations. Biometrika, 64,21-28.

HINKLEY,D.V. (1978). Improving The Jackknife With Special Reference To Correlation Estimation. Biometrika, 65, 13-21.

JONES, H.L. (1974). Jackknife Estimation Of functions Of Stratum Means., Biometrika, 61(2), 343-348.

* KISH, L. (1949). A Procedure For Objectve Respondent Selection Within The Household. Journal Of The American Statistical Association, 44, 380-387.

KISH, L. FRANKEL. M.R. (1974). Inference From Complex Samples. Statist. Soc. Sec., B36, 1-37.

KREWSKI, D. , RAO, J.N.K. (1981). Inference From Stratified Samples: Properties Of The Linearization Jackknife And Balanced Repeated Replication Methods. The Annals Of Statistics, 9, 1010-1019.

* LANDRY, R. (1987). Sélection du répondant lors de sondages téléphoniques, Recueil des textes des présentations du colloque sur les applications de la statistique, (Bureau de la statistique du Québec). 77-95.

LEMESHOW, S., EPP, R. (1977). Properties Of The Balanced Half-Sample And Jackknife Variance Estimation Techniques In The Linear Case. Commun. Statist.-Theo. Meth. A6(13),1259-1274.

* MILLER, R.G. (1974). The jackknife - A Review, Biometrika, 61 (1), 1.

* PANEL ON INCOMPLETE DATA.(1983). Incomplete Data In Sample Surveys. (vol. 1: Report And Case Studies). (William G. Meadow, Ingram Olkin, Donald B. Rubin, éd.). New-York: Academic Press. Inc.

* RAO, J.N.K., WEBSTER, J.T. (1966). On Two Methods Of Bias Reduction In The Estimation Of Ratios, Biometrika, 53, (3-4), 571.

* SOM, R. K. (1976). A Manuel Of Sampling Techniques. London. Heineman.

* STEPHAN, F. F. (1945). The Expected Value And Variance Of The Reciprocal And Other Negative Powers Of A Positive Bernouillian Variate,Annals Of Mathematical Statistics, 16, 50-61.

TEN CATE, A. (1987). Variance Estimation Of The Mean Of a Weighted Sample File. Computational Statistics Quarterly, 4, 197-205.

THOMSEN, I. (1976). Evaluating The Efficiency Of Two Weighting Procedures To Reduce Non-response Bias: An Application Of Cochran's Non-response Stratum Approach. Bull. Inst. Internat. Statist., 46(4), 396-401.

* TUKEY, J. W. (1958). Bias And Confidence In Not-Quite Large Samples. Ann. Math. Statist, 29 (614), 1.

WILLIAMS, W. H. (1962). The Variance Of An Estimator With Post-Stratified Weighting. American Statistical Association Journal ,57,623-627.

WILLIAMS, W.H. (1964). Sample Selection And The Choice Of Estimator In Two-way Stratified Populations.American Statistical Association Journal, December, 1054-1062.

WOODRUFF, R.S. (1971). A sample Method For Approximating The Variance Of A Complicated Estimate. American Statistical Association Journal, 66(334), 411-414.

---

* signifie que la référence est citée dans le texte

UNE APPROCHE INTÉGRÉE DU TRAITEMENT DES VALEURS ABERRANTES DANS LES ENQUÊTES ÉCONOMIQUES INFRA-ANNUELLES

Jean-Louis Tambay, Statistique Canada
4-D8 immeuble Jean Talon, Ottawa (Ontario) K1A 0T6

RÉSUMÉ

Dans les enquêtes économiques infra-annuelles, l'estimation des tendances peut être aussi importante que celle des niveaux. Ce double objectif complique les problèmes de la détection et du traitement des données aberrantes. La présente décrit une étude visant à comparer plusieurs méthodes de traitement des données aberrantes. On s'intéresse à l'effet des traitements sur les estimations et leur variance. Une stratégie qui traite des données aberrantes vis-à-vis les niveaux et les tendances est suggérée.

MOTS CLÉS : Données aberrantes, niveaux, tendances, winsorisation.

1. INTRODUCTION

Les enquêtes économiques infra-annuelles, comme celles menées par Statistique Canada, ont pour but l'estimation des niveaux et des tendances de l'activité économique. Dans bon nombre de ces enquêtes, les échantillons se chevauchent donc considérablement d'une période à l'autre. Ceci complique le problème du traitement des données aberrantes (que nous désignerons par le sigle DA), c.-à-d. les observations vérifiées qui ont une influence forte, mais non-méritée, dans l'estimation. Certaines des méthodes utilisées pour traiter les grandes valeurs observées dans les enquêtes ponctuelles (Hidiroglou et Srinath, 1981) peuvent fausser l'estimation des changements lorsqu'on les applique aux unités communes aux échantillons de deux occasions ou plus.

Afin de définir une méthode de traitement des DA dans les enquêtes infra-annuelles, nous avons comparé empiriquement plusieurs stratégies possibles. Cette communication présente les résultats d'une étude dans laquelle on a mesuré les effets de ces stratégies sur des estimations mensuelles des niveaux et des tendances. Au moyen de ces stratégies, on a traité les DA (de niveau, de tendance ou encore des deux) présentes dans les données d'un échantillon couvrant une période de quatorze mois.

Cette communication comporte cinq sections. La section 2 présente les principaux éléments des stratégies, les variables étudiées, les règles de détection des DA et leurs traitements. La section 3 décrit l'étude empirique et en présente les résultats. Une méthode pour réduire le nombre de DA à traiter est proposée dans la section 4. La section 5 contient quelques remarques générales.

2. MÉTHODOLOGIE

2.1 Cadre général

Les stratégies élaborées pour traiter les DA tiennent compte des types d'enquêtes économiques infra-annuelles pour lesquelles : i) les tendances

et les niveaux présentent un intérêt; ii) l'échantillon est continu, avec un chevauchement considérable pour les périodes voisines; et iii) on peut utiliser une stratification, les estimations étant alors obtenues par application d'un poids d'échantillonnage aux données. Les unités échantillonnées dans une strate n'ont pas nécessairement le même poids. Toutefois, il ne devrait pas y avoir de changement trop prononcé d'une période à l'autre.

2.2 Les valeurs pondérées dans le problème des DA

Soit $\hat{Y} = \Sigma_s$ wy l'estimateur pondéré du total pour une période donnée, où "w" est le poids nominal de l'échantillon, et "y" est la valeur de la variable d'intérêt pour chaque unité. La sommation se fait sur "s", l'ensemble des unités de l'échantillon. Afin d'alléger la notation, on a omis les indices des unités. Si l'estimateur de la période précédente est donné par $\hat{Y}' = \Sigma_{s'}$ w'y', où les apostrophes dénotent l'échantillon, les poids et les valeurs de la période précédente, l'estimation du changement par rapport à la période précédente est alors donnée par

$$\hat{Y} - \hat{Y}' = \Sigma_s \, wy - \Sigma_{s'} \, w'y'$$

$$= \Sigma_c \, w(y-y') + \Sigma_c \, (w-w')y - \Sigma_c \, (w-w')(y-y') + \Sigma_a \, wy - \Sigma_r \, w'y'$$

où "c" représente l'ensemble des unités échantillonnées en commun, "a" représente les unités ajoutées et "r" les unités retirées.

Des trois termes comportant les unités communes de l'échantillon, les deux derniers sont habituellement peu importants par rapport au premier, à moins que les poids ne changent de façon marquée entre les deux périodes. En ce qui concerne les unités non-communes, les valeurs pondérées wy pour les ajouts à l'échantillon peuvent être analysées et traitées pour tenir compte des DA, alors qu'on s'occupe habituellement peu des unités après qu'elles soient retirées de l'échantillon. Pour cette raison, et parce que l'estimateur total utilise les valeurs pondérées wy, on applique la méthode de détéction et de traitement des DA aux valeurs pondérées wy et wd = w(y-y'). Comme nous l'avons mentionné, les poids peuvent ne pas être égaux dans une strate.

2.3 Détéction des DA

Comme les variables économiques suivent habituellement des lois qui sont asymétriques positives, on détecte les DA de niveau en appliquant à wy des seuils à un côté. On détecte les DA de tendance en appliquant à wd des seuils à deux côtés. On utilise la méthode des quartiles parce qu'elle est simple, non-paramétrique et possède un point de rupture élevé (en d'autres mots, cela signifie approximativement que la détection des DA est robuste à la présence de DA dans les données).

Dans chaque strate, les seuils sont calculés à partir des quartiles de wy et wd pour l'échantillon. Si on trie les n unités échantillonnées de la strate en ordre croissant de la valeur de la variable étudiée, la définition utilisée pour le qième quartile est l'observation [q(n+1)/4]. On fait appel à l'interpolation linéaire si l'expression n'est pas un entier. Notons que le deuxième quartile est la médiane de l'échantillon.

Pour les niveaux, le seuil supérieur est :

$$SS = MED(wy) + K1\,[Q3(wy)-MED(wy)];$$

pour les tendances, les seuils supérieur et inférieur sont respectivement :

$$SS = MED(wd) + K2\,[Q3(wd)-MED(wd)], \text{ et } SI = MED(wd) - K2\,[Q3(wd)-MED(wd)],$$

où $Q1(\cdot)$, $MED(\cdot)$ et $Q3(\cdot)$ dénotent les premier, deuxième et troisième quartiles pour le terme de l'argument obtenu de l'échantillon. Cette règle est similaire à la règle dite des "clôtures" ("rule of fences") donnée par Tukey (1977). Les valeurs K1 et K2 sont des constantes utilisées pour étalonner les pourcentages des observations identifiées comme étant des DA.

2.4 Traitement des DA

On envisage deux méthodes pour le traitement des DA: la winsorisation et la méthode de Dalén (1987). Pour les niveaux, la valeur y est remplacée par la valeur y° définie comme suit :

$$\text{Winsorisation : } y^\circ = \begin{cases} y & \text{si } wy \le SS \\ SS/w & \text{sinon;} \end{cases}$$

$$\text{Dalén : } y^\circ = \begin{cases} y & \text{si } wy \le SS \\ (1/w)y + [1-(1/w)](SS/w) = SS/w + (1/w)[y-(SS/w)] & \text{sinon.} \end{cases}$$

Comme on le voit, la winsorisation remplace la valeur aberrante par son seuil, SS/w. Cette approche est différente de la winsorisation habituelle, laquelle est utilisée lorsque les poids sont égaux et qui remplace la valeur aberrante par une valeur limite de l'échantillon. Ici, SS/w ne correspond pas nécessairement à une valeur de l'échantillon.

On peut considérer la méthode de Dalén comme une combinaison linéaire de la valeur originale et de la valeur obtenue par winsorisation. D'autre part, elle réduit à un le poids de [y-(SS/w)], soit la partie de y qui est au-dessus du seuil SS/w.

On peut faire la même chose pour les différences pondérées, wd. Lorsqu'on obtient une valeur de remplacement d°, la valeur réelle y est remplacée par $y^\circ=y'+d^\circ$. On notera que, dans le cas des tendances, on applique sur wd la procédure des deux côtés, et y° peut alors être plus grand ou plus petit que y.

Ce sont là les éléments de base des stratégies visant à traiter les DA. Nous indiquerons les stratégies après la description des données d'enquête utilisées dans l'étude.

3. ÉTUDE EMPIRIQUE

3.1 Description des données utilisées dans l'étude

On a appliqué les stratégies aux données historiques de l'enquête mensuelle de Statistique Canada sur les livraisons, stocks et commandes. Les données sur les livraisons mensuelles de produits fabriqués, entre août

1986 et septembre 1987, ont été tirées d'unités d'échantillon dans quatorze industries.

Cet échantillon est stratifié par taille en trois strates pour chaque industrie et province. L'échantillon initialement choisi dans chaque strate est un échantillon aléatoire simple; toutefois, lorsque les unités qui changent d'industrie sont déplacées dans leur nouvelle strate d'industrie, l'échantillon finit par contenir des unités de poids variés dans une même strate. Aux fins de l'étude, on a traité l'échantillon comme s'il était aléatoire simple stratifié, tiré d'une population fixe pendant la période d'étude. Pour obtenir les nouveaux poids d'échantillon, les comptes ont été réalisés à partir de la base de sondage.

Il y avait 144 strates non-autoreprésentatives dans les industries étudiées. Comme l'indique le tableau ci-dessous, la plupart avait une faible taille d'échantillon. Le traitement des DA a été appliqué seulement aux strates non-autoreprésentatives ayant 7 unités d'échantillon ou plus. Le reste des strates non-autoreprésentatives représentait moins de 3 % de la valeur totale estimée des livraisons.

taille d'échantillon	nombre de strates	variation des poids dans les strates	
		minimum	maximum
2 - 6	56	1,167	42,50
7 - 14	43	1,125	23,73
15 - 22	21	1,190	17,25
23 - 30	14	1,071	13,68
plus de 30	10	1,136	9,86

3.2 Stratégies étudiées pour le traitement des DA

On a étudié 14 stratégies de traitement des DA. Ces stratégies sont caractérisées par la méthode utilisée pour remplacer les DA (winsorisation ou méthode de Dalén), par la façon dont on les a appliquées aux DA de niveau et de tendance (stratégies des niveaux seulement, stratégies des tendances seulement, stratégies combinées, stratégies intégrées), et par le pourcentage global de valeurs pondérées, wy et wd, identifiées comme étant des DA (2 % et 5 %).

On a obtenu des taux de DA de 2 % en initialisant K1 à 5,99 et K2 à 20,70 dans les équations de seuil. Avec K1 = 3,38 et K2 = 9,80, on a obtenu des taux de 5 %.

Dans l'approche "niveaux seulement", on traite uniquement les DA de niveau (wy), et dans l'approche "tendances seulement", on traite uniquement les DA de tendance (wd), comme on l'a expliqué à la section 2. Pour les DA de tendance, on a appliqué cette procédure d'abord aux données les plus anciennes (p. ex., les différences entre août et septembre 1986), et les valeurs remplacées ont eu un effet sur les tendances futures, comme nous le voulions. Les deux approches ont seulement remplacé les valeurs courantes pour les mois examinés.

L'approche "combinée" consiste à traiter d'abord les DA de tendance pour chaque mois, et ensuite à appliquer la procédure pour les DA de niveau aux données traitées. Si le remplacement de la valeur d'une DA de niveau

faisait de celle-ci une DA de tendance, le remplacement n'était pas effectué pour cette valeur.

L'approche "intégrée" consiste à traiter les données pour le mois précédant la période d'étude (septembre 1986), afin de déceler les DA de niveau. Si une valeur a été changée, alors toutes les autres valeurs du mois pour l'unité ont été changées d'un même pourcentage. On a ensuite appliqué aux données la procédure de traitement des DA de tendance. Cette procédure a simulé les effets du traitement de nouvelles unités d'échantillon qui sont des DA de niveau, en diminuant leur contribution ultérieure aux estimations.

On a utilisé la winsorisation pour les approches suivantes : niveaux seulement, tendances seulement, et approche combinée. On a utilisé la méthode de Dalén pour les quatre approches. Ces 7 traitements, appliqués aux taux de DA de 2 % et 5 %, ont donc constiué les 14 stratégies de traitement des DA.

Les 14 mois de données ont fourni 14 ensembles d'estimations de niveau et 13 ensembles de tendances. Toutefois, on n'a étudié que les 12 derniers ensembles d'estimations de niveau et tendance. Les autres mois ont servi à "démarrer" les procédures de traitement des DA de tendance.

3.3 Résultats de l'étude

L'étude s'est attardée aux effets des traitements des DA sur les estimations mensuelles obtenues au niveau de l'industrie. Les estimations des niveaux ($\hat{Y}$) et des changements mensuels ($\hat{Y} - \hat{Y}'$), et leurs variances, ont été calculées au moyen des équations standards pour les échantillons aléatoires simples stratifiés.

On a calculé les estimations et les variances pour les données traitées en substituant les données de remplacement aux données originales dans les équations standards.

On a ainsi obtenu trois mesures relatives des effets des traitements des DA. Elles mesurent le biais relatif des traitements, la diminution de la variance due au traitement, et l'erreur quadratique moyenne pour les données traitées par rapport aux données non traitées. Ces mesures sont les suivantes :

Biais relatif estimé = [Est(Trmt) - Est(Orig)] / Errtyp(Orig)
Erreur type relative = Errtyp(Trmt) / Errtyp(Orig)
Erreur quad. moyenne relative = (Biais rel. est.)^2 + (Err. type rel.)^2.

Les figures 1 et 2 illustrent les graphes des deux premières mesures, pour les estimations industrielles mensuelles, des niveaux (totaux) et des tendances (changements d'un mois à l'autre), respectivement. Les graphes représentent les résultats des 7 stratégies de traitement qui utilisent un taux de 2 % pour les DA. Les stratégies sont identifiées par deux caractères (WC pour winsorisation combinée, etc.) et deux chiffres (02 pour les taux de DA de 2 %), sur chaque graphe. Chaque graphe comporte 168 observations, correspondant à douze estimations mensuelles pour chacune des quatorze industries.

Les axes des graphes ne sont pas tous identiques. Afin de faciliter les comparaisons, on a superposé des rectangles standards sur chaque graphe. Le haut et le bas de chaque rectangle indiquent les points où les traitements n'ont pas eu d'effet sur les erreurs types, et où ils les ont réduits de moitié. Les limites latérales indiquent les points où les biais sont égaux à la moitié des erreurs types estimées originalement, ce qui n'est pas une quantité négligeable.

On a également tracé le cercle unitaire sur chaque graphe, car il montre les points correspondant à une erreur quadratique moyenne (EQM) relative égale à un. Les points à l'extérieur du cercle représentent les pertes de "rendement" en raison du traitement. Les points à l'intérieur représentent les gains.

Les observations suivantes s'imposent :

- La winsorisation influence fortement bon nombre des estimations, ce qui produit de nombreuses pertes de rendement.

- Les méthodes de Dalén offrent un meilleur résultat à cet égard.

- Pour les estimations de niveau, la plupart des biais sont dans la même direction. Les erreurs types augmentent rarement.

- Les stratégies DN02, DC02 et DI02 semblent avoir eu des effets similaires sur les estimations de niveau. La stratégie DT02 a eu un plus faible effet, ce qui n'était pas tout à fait surprenant.

- En ce qui concerne les tendances, les méthodes DT02 et DI02 ont eu des effets similaires. DN02 a eu un effet plus faible sur les tendances, mais il y a eu un plus grand nombre de pertes de rendement avec DC02.

La figure 3 illustre certains des graphes obtenus avec un taux de DA de 5 %. Les effets des traitements sur les estimations sont davantage prononcés, et il y a plus de pertes de rendement. Il y a une exception, soit la méthode DT pour les tendances, laquelle n'est pas autant perturbée par le nombre de DA traitées. Cela est probablement dû au fait que la méthode est appliquée symétriquement aux données.

Le tableau ci-dessous donne un résumé des graphes. Dans le tableau, on retrouve, pour chaque méthode, la moyenne des 168 EQM relatives calculées, ainsi que les moyennes des EQM relatives calculées au niveau de l'ensemble des industries. Ces moyennes sont établies pour les douze mois de l'étude.

Les moyennes industrielles confirment les observations faites précédement : pour les taux de DA étudiés, la winsorisation n'est pas aussi efficace que la méthode de Dalén; de plus, la méthode DT02 a un plus faible effet sur les estimations de niveau que les autres méthodes de Dalén - et son effet est près de celui de la stratégie DT05; enfin, les autres traitements ne sont pas aussi efficaces lorsque le taux de DFA est de 5 %.

Moyennes mensuelles des EQM relatives

Stratégie de traitement	Niveaux		Tendances	
	Moyenne industrielle	Moyenne globale	Moyenne industrielle	Moyenne globale
WN02	1,35	2,28	1,38	1,10
WT02	1,17	0,99	1,20	0,92
WC02	1,21	1,35	1,42	0,99
DN02	0,95	1,32	1,03	0,95
DT02	0,99	0,98	0,97	0,86
DC02	0,95	1,20	1,08	0,84
DI02	0,98	1,38	0,96	0,87
WN05	1,92	6,21	1,54	1,13
WT05	1,26	1,05	1,29	0,78
WC05	1,40	2,10	1,54	0,91
DN05	1,04	3,02	1,04	0,88
DT05	0,99	1,00	0,93	0,62
DC05	1,01	1,69	1,07	0,80
DI05	1,06	2,81	0,91	0,65

Il y a lieu de noter la différence entre les moyennes industrielles et les moyennes globales des EQM relatives. Pour les niveaux, cette différence est grande et positive pour les approches de traitement des DA de niveau. Cela s'explique par le fait que, à la différence de l'approche "tendances seulement", le biais global (ensemble des industries) est un regroupement de biais industriels qui sont presque tous dans la même direction. En fait, les résultats globaux pour les niveaux ne semblent pas encourager l'utilisation de quelque méthode de traitement que ce soit.

En ce qui concerne les tendances, ce sont les moyennes globales qui sont les plus basses. Cette réduction est due en grande partie à la composante "erreur type relative" de l'EQM relative. Elle diminue de façon appréciable au niveau global. Le fait que toutes les méthodes présentent des biais pour les tendances dans les deux directions contribue également à la valeur plus faible des EQM relatives globales. Les biais ont tendance à s'annuler plus souvent, ce qui produit un faible biais global.

4. RÉDUCTION DES DA À TRAITER

La détection et le traitement des DA sont faits indépendamment dans chaque strate en se servant de seuils obtenus de l'échantillon dans la strate. L'importance inégale des strates à l'estimation des niveaux industriels suggère d'examiner de plus près l'effet du traitement des DA sur les estimations industrielles. Ceci nous mène à proposer une méthode pour réduire le nombre de DA à traiter.

Soit $w|y-y^{\circ}|/\hat{Y}$ l'effet du traitement d'une DA, où "y" est la valeur originale, "y°" est la valeur de remplacement, "w" est le poids nominal de l'échantillon et "Y" est l'estimation du niveau industriel basée sur les données originales pour le même mois. La valeur absolue de l'effet est utilisée. Le tableau ci-dessous donne les fréquences observées des effets du traitement des DA pour six des stratégies à l'étude. On note que le nombre de DA traitées selon l'approche "combinée" est supérieur à celui des deux autres méthodes car l'approche traite les DA de niveau et de tendance.

Effet du traitement des DA sur l'estimation des niveaux industriels

Stratégie de traitement	Nombre total de DA traitées	Fréquence des effets de traitement moins de 0,1 %	0,1 à 0,5 %	0,5 % et plus
DN05	950	410	300	240
DT05	985	615	270	100
DC05	1555	865	480	210
DN02	390	140	140	110
DT02	390	240	100	50
DC02	680	340	210	130

Pour chaque stratégie, moins de 30 % des DA ont un effet de traitement de 0,5 % ou plus. Ceci veut dire qu'une grande majorité des DA ont subi un traitement qui est relativement insignifiant dans leur industrie. On a ainsi une règle pour réduire le nombre de DA à traiter, soit en ne traitant que les DA dont l'effet du traitement à un niveau donné (par ex. l'industrie) est inférieur à un seuil prédéterminé. Si jamais on se servait de cette règle avec un seuil de 0,5 % pour la stratégie DC02, le nombre moyen de DA à traiter par industrie tomberait de 4.0 à 0.8 par mois.

5. OBSERVATIONS ET CONCLUSIONS

L'étude a démontré les possibilités offertes par les diverses approches de traitement des DA de niveau et de tendance dans les enquêtes infra-annuelles, et il convient de les analyser plus à fond. La méthode choisie pour combiner les deux dépend des facteurs comme l'importance relative des estimations de niveau et de changement, ou des regroupements pour les niveaux industriel et global, et évidemment des facteurs opérationnels. Il pourrait s'avérer plus difficile d'intégerer dans une enquête les procédures de traitement des DA de tendance.

Tout comme les méthodes se complètent, chacune comporte des risques. Par exemple, une valeur très inhabituelle présente dans un mois donné peut avoir des répercussions pendant plusieurs mois à cause de la procédure de traitement des DA de tendance qui décale les valeurs futures dans la direction de cette donnée aberrante. Il sera toujours nécessaire de bien détecter les DA et d'établir des garde-fous.

REMERCIEMENTS

L'auteur tient à remercier Jack Gambino, Hyunshik Lee et Claude Poirier de Statistique Canada pour leurs commentaires et suggestions forts utiles.

BIBLIOGRAPHIE

Dalén, J. (1987). Practical Estimators of a Population Total Which Reduce the Impact of Large Observations. Rapport technique. Statistics Sweden.

Hidiroglou, M.H. et Srinath, K.P. (1981). Some Estimators of a Population Total From Simple Random Samples Containing Large Units. **Journal of the American Statistical Association**, 76, 690-695.

Tukey, J.W. (1977). **Exploratory Data Analysis.** Reading, MA : Addison-Wesley.

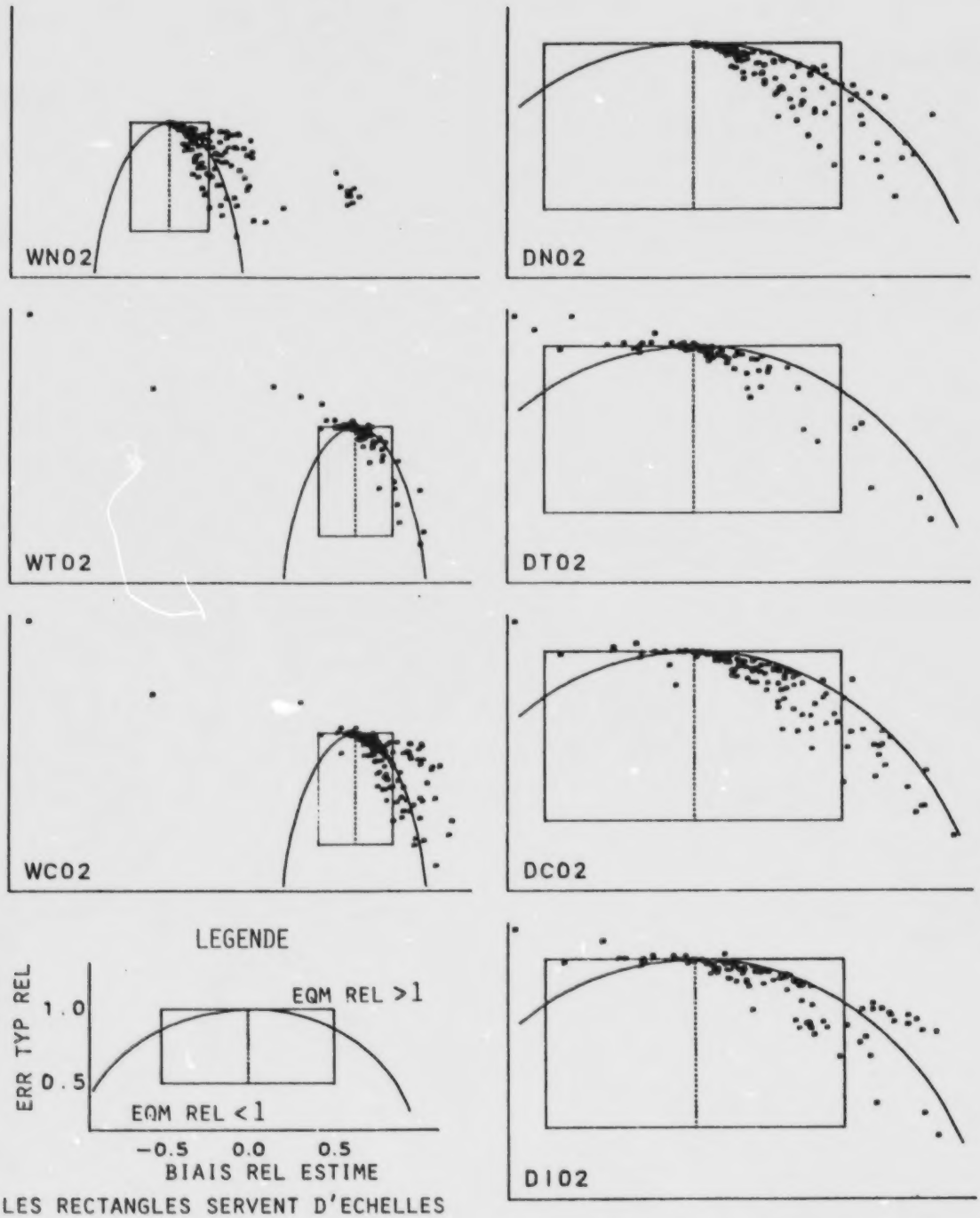


FIGURE 1 - EFFET DES TRAITEMENTS SUR LES ESTIMATIONS MENSUELLES DES NIVEAUX.

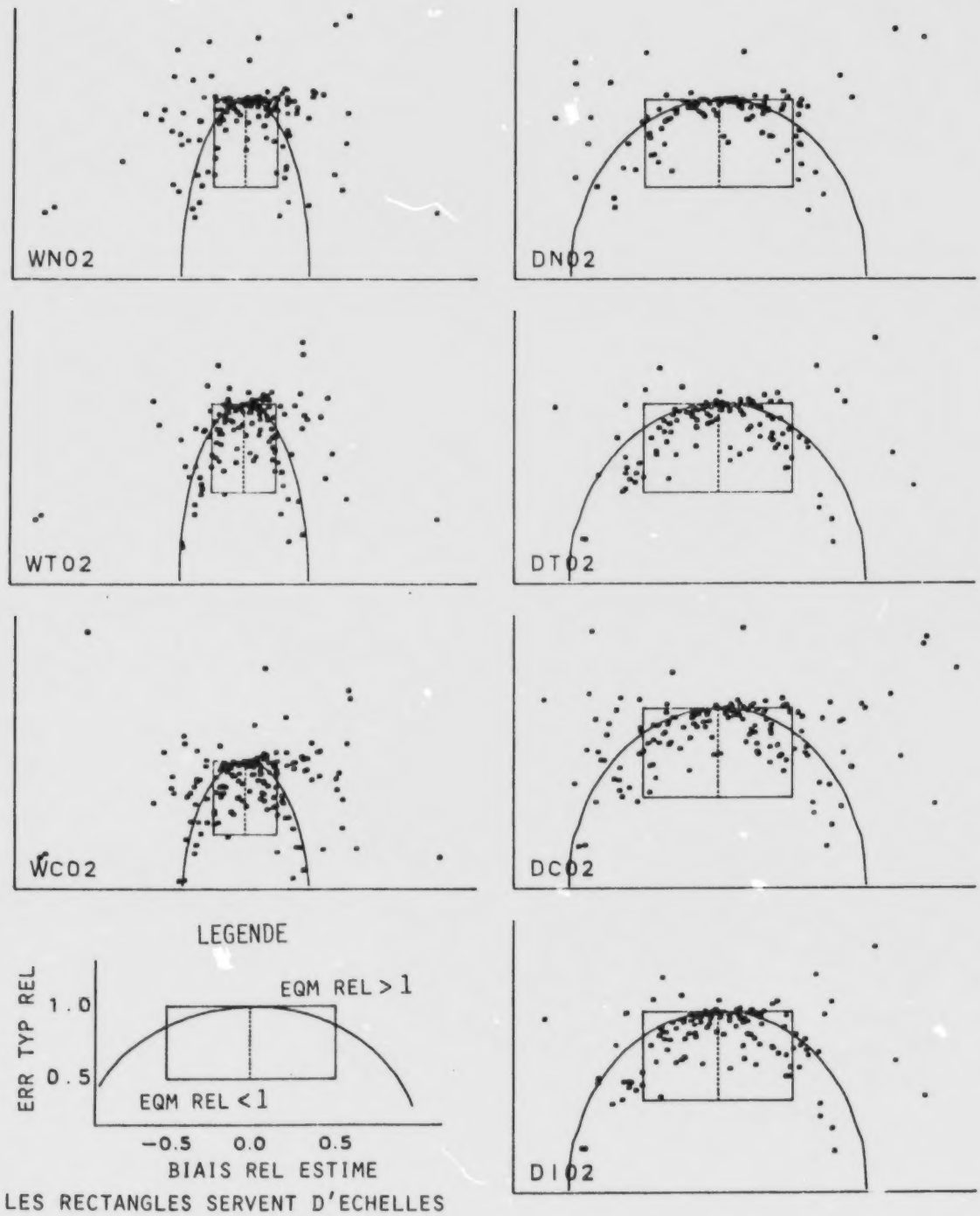


FIGURE 2 - EFFET DES TRAITEMENTS SUR LES ESTIMATIONS MENSUELLES DES TENDANCES.

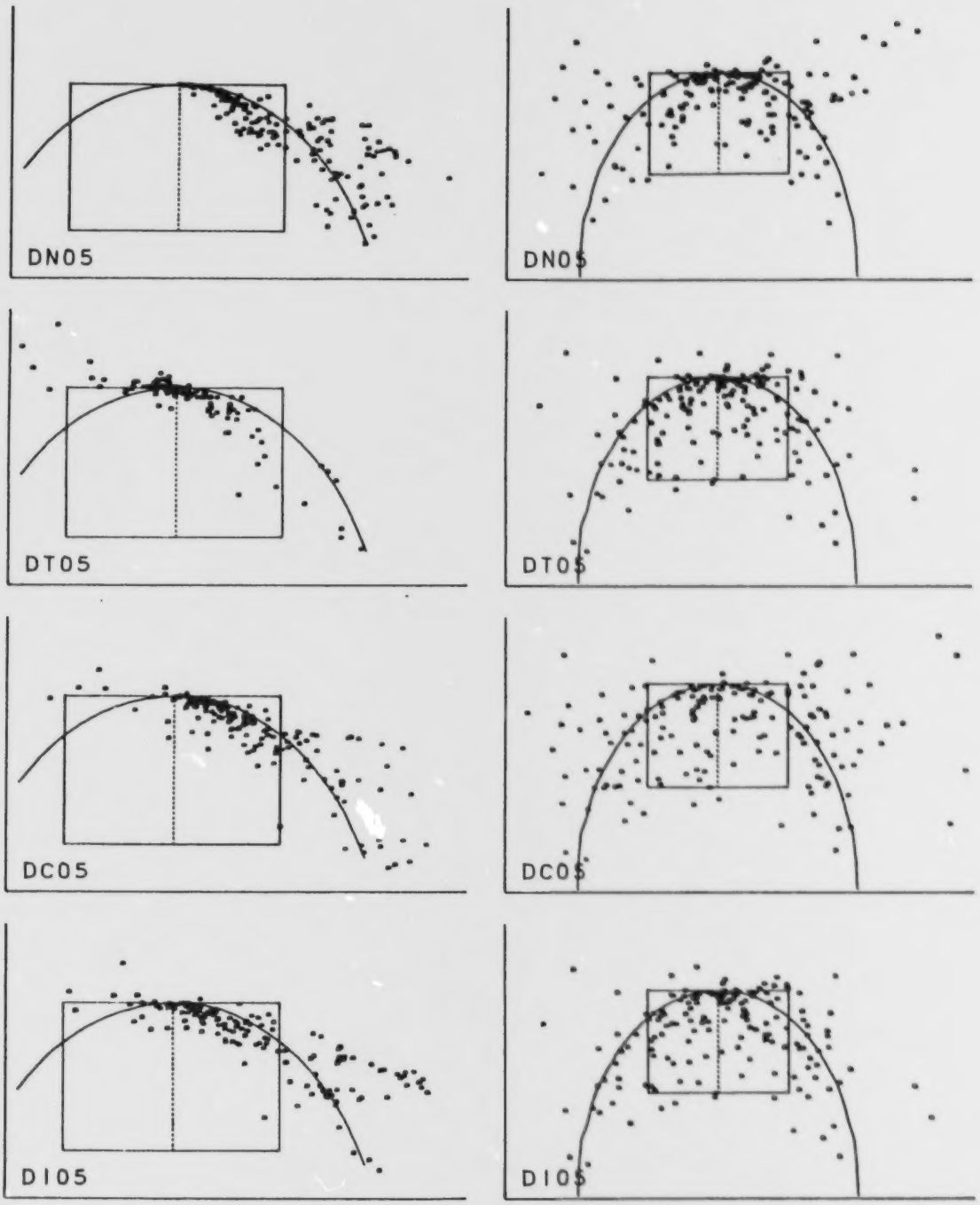


FIGURE 3- EFFET DES TRAITEMENTS SUR LES ESTIMATIONS MENSUELLES DES NIVEAUX (A GAUCHE) ET DES TENDANCES (A DROITE).

PROBLEMES D'ECHANTILLONNAGE RENCONTRES DANS L'ADMINISTRATION DE PROGRAMMES D'ASSURANCE-RECOLTE

par
Sylvain Végiard, statisticien
Régie des assurances agricoles du Québec
Direction des services techniques et de la recherche
Le 18 mai 1989

Résumé

L'expertise en assurance-récolte consiste à estimer le rendement obtenu au terme de la saison de production par un groupe d'assurés (système collectif) ou chaque assuré (système individuel). Les techniques d'échantillonnage sont utilisées à cette fin dans plusieurs programmes d'assurance.

Le plan d'échantillonnage généralement utilisé dans le système individuel est illustré en considérant l'exemple du maïs-grain. Un plan d'échantillonnage à trois degrés est proposé pour le système collectif (catégorie foin). Les formules des estimateurs et de leurs variances sont présentées.

1. L'organisme et ses produits

La Régie des assurances agricoles du Québec a été mise sur pied en 1981, à la faveur de la fusion de deux organismes : la Régie de l'assurance-récolte et la Commission administrative des régimes d'assurance-stabilisation des revenus agricoles. Son siège social est situé à Lévis, et un réseau de treize bureaux régionaux vient appuyer ses opérations dans les principaux secteurs agricoles du Québec.

La Régie offre aux producteurs agricoles deux formes de produits. L'assurance-stabilisation des revenus agricoles protège contre les fluctuations de prix des marchés, tandis que l'assurance-récolte couvre les pertes subies suite à des causes naturelles contre lesquelles il n'existe pas de protection adéquate (climat défavorable, épidémies d'insectes ou de maladies, etc...).

1.1 Effets recherchés par l'assurance agricole

En période difficile (économique ou climatique), l'assurance agricole peut éviter la faillite d'entreprises qui doivent continuer à remplir leurs obligations financières en dépit d'une perte de revenus qui échappe à leur contrôle. A l'échelle québécoise, l'assurance peut empêcher que certaines productions soient abandonnées uniquement parce qu'elles ont à traverser une période momentanément défavorable, contribuant ainsi à préserver l'intégrité du climat socio-économique.

Un autre avantage de l'assurance réside dans la stabilisation des dépenses gouvernementales. En effet, les trésors publics préfèrent débourser annuellement une montant relativement fixe (les gouvernements fédéral et provincial partagent le financement des fonds d'assurance avec les producteurs assurés), plutôt que d'effectuer des paiements «ad hoc» lors d'années catastrophiques.

1.2 Fonctionnement de l'assurance-récolte

L'assurance-récolte est disponible selon le système individuel ou le système collectif, suivant les diverses productions végétales qui sont assurées par la Régie.

Dans le système individuel, le producteur s'assure contre le risque de voir son propre rendement annuel être en deçà de 80% du rendement moyen qu'il a lui-même obtenu lors des années précédentes.

Dans le système collectif, le territoire québécois est découpé en un certain nombre de zones d'assurance, et le producteur s'assure contre le risque de voir le rendement annuel de sa zone être en deçà de 80% du rendement moyen que sa zone a obtenu lors des années précédentes.

Si, au moment de la récolte, le rendement annuel s'avère inférieur à 80% du rendement moyen, des indemnités seront versées au producteur pour combler monétairement cet écart.

1.3 Méthodes d'estimation des rendement annuels (expertise)

Afin d'établir le rendement obtenu au terme de la saison de végétation, la Régie procède à ce que l'on appelle une expertise. Trois méthodes sont principalement utilisées pour effectuer cette opération :

1.3.1 Echantillonnage

Cette méthode consiste à prélever dans le champ un échantillon de matière végétale au moment où le producteur effectue la récolte. Ceci implique donc la réalisation des étapes suivantes:

(i) Découpage de la population

Dans le système collectif, le territoire est tout d'abord fractionné en zones (147 pour le programme Grandes cultures en 1989). L'intérêt porte sur l'estimation du rendement annuel pour chacune de ces zones. Les producteurs de chaque zone constituent les unités primaires d'échantillonnage, et les champs de chaque producteur forment les unités secondaires. Finalement, des sites (généralement de surface d'un mètre carré) seront sélectionnés dans les champs faisant partie de l'échantillon.

Dans le système individuel, l'intérêt est mis sur l'estimation du rendement annuel pour chaque producteur. Les champs du producteur constituent les strates, et des sites (de surface variant généralement de un à trois mètres carrés) seront choisis dans chacun des champs du producteur.

(ii) Sélection d'un échantillon probabiliste

La sélection des unités d'échantillonnage se fait de façon probabiliste, ce qui assure un caractère objectif à la démarche, et permet l'obtention d'estimateurs ayant des caractéristiques désirables, notamment l'absence de biais et la possibilité de mesurer l'erreur d'échantillonnage.

(iii) Cueillette des données sur le terrain

Au moment où le producteur décide d'effectuer sa récolte, un représentant de la Régie se déplace pour prélever un échantillon de matière végétale.

(iv) Traitement des données au laboratoire

Les opérations les plus couramment effectuées sont le séchage des échantillons (les résultats sont souvent exprimés à un taux d'humidité standardisé) et la pesée.

(v) Estimation

Cette étape consiste à regrouper les données provenant des diverses unités d'échantillonnage pour obtenir le rendement annuel de zone (dans le cas du système collectif) ou le rendement annuel du producteur (dans le cas du système individuel).

Les coûts reliés à cette méthode sont élevés, et nécessitent un effort logistique important. De plus, la taille d'échantillon à prélever pour atteindre un niveau de fiabilité acceptable dépend du niveau de variation enregistré, qui à son tour, fluctue selon les années. Il est donc difficile de prescrire une taille d'échantillon qui donnera satisfaction à chaque année, en raison du caractère imprévisible de la variance entre les unités d'échantillonnage.

Cette méthode possède toutefois l'avantage de permettre la mesure de l'erreur d'échantillonnage, qui renseigne l'utilisateur sur la qualité de l'estimation.

1.3.2 Décompte physique

Pour certaines cultures, le producteur doit forcément faire séjourner sa récolte dans un entrepôt ou un silo. Il est donc possible pour un représentant de la Régie d'estimer sur place le rendement obtenu lorsque le producteur a complété ses opérations de récolte.

Par exemple, dans le cas de céréales ou de maïs-grain stockés dans un silo, on peut estimer le volume et la densité du grain, et en déduire une quantité récoltée.

Toutefois, la mesure du volume de la récolte est complexe, car bien que le silo soit cylindrique, le grain s'y trouvant à l'extrémité supérieure forme un cône irrégulier en raison du mode de emplissage. Il découle que certaines approximations doivent être faites à ce niveau.

De plus, la densité des grains n'est pas uniforme sur toute la hauteur du silo, parce que plusieurs cargaisons de camion sont nécessaires pour en effectuer le remplissage, et que chaque cargaison recèle potentiellement une densité de grains différente. Or, seuls les grains de la portion supérieure du silo sont accessibles pour mesurer la densité de tous les grains du silo. Il s'ensuit que l'échantillon prélevé peut difficilement être considéré comme représentatif du contenu total du silo, et que l'erreur d'estimation ainsi commise est difficilement quantifiable.

1.3.3 Preuves de vente

Dans les cas où le producteur vend sa récolte en totalité à des acheteurs (centres régionaux des grains, usines de congélation ou de transformation), on peut envisager l'utilisation des factures émises au moment de la transaction comme source de données.

Cependant, cette approche ne peut être étendue à toutes les situations, notamment lorsque le producteur ne commercialise qu'une partie de sa récolte avec des preuves de vente, consommant le reste sur sa ferme ou en le vendant en kiosque.

1.4 Option privilégiée à la Régie

Bien que les trois techniques comportent des désavantages, l'approche faisant appel à l'échantillonnage possède certains attraits (objectivité, capacité de mesurer l'erreur d'échantillonnage, possibilité d'application à presque toutes les situations) qui font en sorte qu'on tentera, dans la mesure du possible, de la privilégier dans les opérations d'expertise effectuées en assurance-récolte.

2. Le système individuel : l'exemple du maïs-grain

Le but de l'expertise est d'obtenir une estimation du rendement annuel du producteur (RP), exprimée en kilos à l'hectare, et une mesure de fiabilité de cette estimation.

Un producteur peut avoir plusieurs champs de maïs-grain, disons N_i, chacun ayant un rendement en kilos à l'hectare, noté RC_i. Un champ est considéré ici comme une strate, de superficie égale à H_i mètres carrés. On prélève dans chaque champ un certain nombre de sites ayant tous une superficie de 3 mètres carrés.

2.1 Estimateurs et variances

Les formule d'estimation sont celles généralement utilisées pour des échantillons aléatoires stratifiés. Le lecteur désireux d'obtenir plus de détails à ce sujet est invité à consulter l'ouvrage de Cochran (1977, chapitre 5).

2.2 Difficultés rencontrées

Comme mentionné à la section 1.3.1, il est difficile d'obtenir à chaque année des estimations qui rencontrent les critères de fiabilité, en raison du caractère imprévisible de la variance intra-champ. Voilà pourquoi la norme actuellement en vigueur (un site à l'hectare, avec un minimum de cinq sites par champ) prévoit un taux d'échantillonnage qui est indépendant du niveau de variation dans le champ.

De plus, l'estimation par le quotient ou la régression est impraticable en raison de l'absence de variables reliées au rendement et disponibles au niveau de chaque site du champ.

2.3 Solutions envisagées

La mesure de l'erreur d'échantillonnage permet d'apprécier la fiabilité d'une estimation au moment de son utilisation. Le contrôle de l'erreur d'échantillonnage implique un niveau d'analyse plus élevé, c'est-à dire, la prescription d'une taille d'échantillon assurant, avec un certain niveau de probabilité, l'obtention d'une erreur inférieure à un seuil de tolérance fixé à l'avance. Dans le cas du maïs-grain, il faudra vraisemblablement se satisfaire de la mesure de l'erreur, compte tenu de l'incertitude entourant la variation intra-champ pour une année donnée.

Le travail qui resterait à faire viserait donc à l'amélioration de la norme actuelle concernant la taille d'échantillon. On pourrait songer à chercher une allocation qui maximise le nombre de producteurs ayant une estimation plus fiable que le seuil de tolérance fixé, avec un effort d'échantillonnage global donné. Les paramètres à faire varier pourraient être les minimum et maximum par producteur, les minimum et maximum par champ, et le nombre de sites à l'hectare.

3. Le système collectif : l'exemple du foin

Le foin sert à l'alimentation des herbivores (vaches laitières, vaches de boucherie, etc...). La production laitière étant l'activité agricole prédominante dans la province, il est facile de comprendre le rôle majeur qu'occupe la production du foin dans l'ensemble du secteur agricole québécois.

La culture du foin comporte certaines caractéristiques qui la distinguent par rapport aux autres productions végétales. Le foin se compose d'une ou plusieurs espèces de plantes, par exemple, des graminées telles que le mil ou le brome, et des légumineuses telles que la luzerne ou le trèfle. Par opposition aux cultures annuelles, le foin est une culture pérenne, c'est-à-dire qu'un champ peut produire plusieurs années de suite sans que l'on y fasse de semis. Selon les régions, on peut récolter une, deux, ou trois coupes de foin par année sur le même champ.

3.1 Projet-pilote pour 1989 : nouveau plan d'échantillonnage à trois degrés

A titre expérimental, un projet-pilote est entrepris cette année dans 4 zones afin d'évaluer un nouveau plan d'échantillonnage conçu pour estimer le rendement annuel de la zone en kilos de foin à l'hectare, et obtenir une mesure de fiabilité sur cette estimation.

Le plan d'échantillonnage comporte trois degrés. Les producteurs dans la zone représentent les unités primaires d'échantillonnage, les champs du producteur forment les unités secondaires, et les sites dans le champ constituent les unités tertiaires.

3.1.1 Sélection des producteurs

Pour chaque producteur de la zone, on connaît la taille de son troupeau exprimée en unités animales. On détermine tout d'abord un certain nombre de producteurs qui seront autoreprésentatifs (AR). Un producteur est jugé autoreprésentatif s'il possède plus de 2.75 fois le nombre moyen d'unités animales de tous les producteurs de la zone. Toutefois, il y aura un maximum de 5 producteurs autoreprésentatifs par zone (dans le cas où plus de 5 producteurs rencontreront ce critère, ce seront les 5 producteurs ayant le plus d'unités animales qui seront désignés AR). Le nombre de producteurs AR dans une zone variera donc de 0 à 5.

Dans un deuxième temps, on identifiera les petits producteurs (PP). Un producteur est déclaré PP s'il a moins de 15 unités animales.

Finalement, le reste des producteurs de la zone sera divisé en 2 groupes : les assurés (AS) à la catégorie foin du programme Grandes cultures selon le système collectif et les non-assurés (NA).

Au total, la zone sera ainsi découpée en 4 groupes mutuellement exclusifs :

- AR : autoreprésentatifs
- PP : petits producteurs
- AS : assurés
- NA : non-assurés

Dans chacun de ces 4 groupes, un échantillon de producteurs sera sélectionné. On aura :

- pour les AR : tous les producteurs sont retenus
- pour les PP : on choisit 2 producteurs seulement
- pour les AS et les NA : on choisit un nombre de producteurs qui respecte les proportions de AS et de NA observées dans la population, et de sorte que le nombre total de producteurs sélectionnés dans la zone soit égal à 25.

3.1.2 Sélection des champs

On connaît la superficie de chaque champ, telle que déclarée par le producteur. On détermine tout d'abord s'il y aura ou non un champ autoreprésentatif (AR) chez ce producteur. Un champ est jugé autoreprésentatif s'il possède une superficie excédant 2.75 fois la superficie moyenne des champs de foin du producteur. Par contre, il y aura un maximum de 1 champ autoreprésentatif par producteur (dans le cas où plusieurs champs répondent à ce critère, ce sera le plus grand champ du producteur qui sera désigné AR). Tous les autres champs du producteur seront désignés AC.

Les champs du producteur seront ainsi séparés en 2 groupes :

- AR : autoreprésentatif
- AC : autres champs

Le nombre de champs à échantillonner dans chacun de ces 2 groupes est établi comme suit:

- pour le AR : il est automatiquement retenu dans l'échantillon
- pour les AC : si le producteur est AR et qu'il n'a pas de champ AR, on choisit au hasard 3 champs. Sinon, on choisit au hasard 2 champs.

3.1.3 Sélection des sites

Nombre de sites à prélever par champ :

- pour le champ AR d'un producteur AR :	6 sites
- pour le champ AR d'un producteur PP, AS ou NA :	4 sites
- pour un champ AC d'un producteur AR :	4 sites
- pour un champ AC d'un producteur PP, AS, ou NA :	2 sites

Les sites sont tous de superficie égale à un mètre carré.

3.1.4 Estimation du rendement du champ, du producteur et de la zone

Dans la notation qui sera employée,les indices prendront les significations suivantes :

i : strate de producteurs (AR, PP, AS ou NA)
j : producteur dans la strate de producteurs
k : strate de champs dans le producteur (AR ou AC)
l : champ dans la strate de champs
m : site dans le champ

Afin d'alléger l'écriture, nous ne considérerons que le cas où une seule coupe de foin a eu lieu dans le champ.

Une estimation du rendement du champ (RC_{ijkl}) peut être obtenue par :

$$\widehat{RC}_{ijkl} = \frac{\sum_{m=1}^{h_{ijkl}} KG_{ijklm}}{h_{ijkl}}$$

où : KG_{ijkl} = la masse en kilogrammes du site m prélevé dans le champ l de la strate k du producteur j de la strate i

h_{ijkl} = le nombre de sites prélevés dans le champ

Puisque la surface des sites est d'un mètre carré, cette estimation est donnée en kilogrammes par mètre carré. Une simple multiplication par un facteur de 10 000 permettrait d'obtenir un rendement en kilogrammes à l'hectare.

Le rendement du producteur peut être estimé par :

$$\widehat{RP}_{ij} = \frac{\sum_{k=1}^{K_{ij}} \frac{C_{ijk}}{c_{ijk}} \sum_{l=1}^{c_{ijk}} H_{ijkl}\, \widehat{RC}_{ijkl}}{\sum_{k=1}^{K_{ij}} \frac{C_{ijk}}{c_{ijk}} \sum_{l=1}^{c_{ijk}} H_{ijkl}}$$

où : H_{ijkl} = la superficie en mètres carrés du champ l de la strate k du producteur j de la strate i

c_{ijk} = le nombre de champs sélectionnés dans la strate k du producteur j de la strate i

C_{ijk} = le nombre total de champs dans la strate k du producteur j de la strate i

K_{ij} = nombre de strates de champs pour le producteur j de la strate i

Finalement, l'estimation du rendement de la zone s'écrit :

$$\widehat{RZ} = \frac{\sum_{i=1}^{I} \frac{P_i}{p_i} \sum_{j=1}^{p_i} \widehat{H}_{ij}\, \widehat{RP}_{ij}}{\sum_{i=1}^{I} \frac{P_i}{p_i} \sum_{j=1}^{p_i} \widehat{H}_{ij}}$$

où : p_i = nombre de producteurs sélectionnés dans la strate i

P_i = nombre total de producteurs dans la strate i

I = nombre de strates de producteurs dans la zone

$$\hat{H}_{ij} = \sum_{k=1}^{K_{ij}} \frac{C_{ijk}}{\mathcal{L}_{ijk}} \sum_{\ell=1}^{\mathcal{L}_{ijk}} H_{ijk\ell}$$

3.1.5 Variance de l'estimateur du rendement de la zone

Cette formule de variance a été obtenue en généralisant l'approche de Cochran (1977, pp. 303-305) décrite pour un échantillon à deux degrés. L'estimation de la variance de l'estimateur du rendement annuel de la zone se présente donc comme suit :

$$\hat{V}(\hat{R}\bar{Z}) = \frac{1}{\hat{H}} \left\{ \sum_{i=1}^{I} \frac{P_i^2(1-p_i/P_i)}{p_i} \frac{\sum_{j=1}^{p_i} (\hat{KG}_{ij} - \hat{H}_{ij}\hat{R}\bar{Z} - \hat{\bar{KG}}_i + \hat{H}_i\hat{R}\bar{Z}/P_i)^2}{p_i - 1} + \right.$$

$$\sum_{i=1}^{I} \sum_{j=1}^{p_i} \frac{P_i}{p_i} \sum_{k=1}^{K_{ij}} \frac{C_{ijk}^2(1-\mathcal{L}_{ijk}/C_{ijk})}{\mathcal{L}_{ijk}} \frac{\sum_{\ell=1}^{\mathcal{L}_{ijk}} (\hat{KG}_{ijk\ell} - H_{ijk\ell}\hat{R}\bar{Z} - \hat{\bar{KG}}_{ijk} + \hat{H}_{ijk}\hat{R}\bar{Z}/C_{ijk})^2}{\mathcal{L}_{ijk} - 1} +$$

$$\left. \sum_{i=1}^{I} \sum_{j=1}^{p_i} \frac{P_i}{p_i} \sum_{k=1}^{K_{ij}} \sum_{\ell=1}^{\mathcal{L}_{ijk}} \frac{C_{ijk}}{\mathcal{L}_{ijk}} \frac{H_{ijk\ell}^2(1-R_{ijk\ell}/H_{ijk\ell})}{R_{ijk\ell}} \frac{\sum_{m=1}^{R_{ijk\ell}} (KG_{ijk\ell m} - \hat{RC}_{ijk\ell})^2}{R_{ijk\ell} - 1} \right\}$$

où :

$$\hat{H} = \sum_{i=1}^{I} \sum_{j=1}^{p_i} \frac{P_i}{p_i} \hat{H}_{ij}$$

$$\hat{KG}_{ij} = \sum_{k=1}^{K_{ij}} \frac{C_{ijk}}{\mathcal{L}_{ijk}} \sum_{\ell=1}^{\mathcal{L}_{ijk}} \frac{H_{ijk\ell}}{R_{ijk\ell}} \sum_{m=1}^{R_{ijk\ell}} KG_{ijk\ell m}$$

$$\hat{\bar{KG}}_i = \sum_{j=1}^{p_i} \hat{KG}_{ij} / p_i$$

$$\hat{H}_i = \sum_{j=1}^{p_i} \frac{P_i}{p_i} \hat{H}_{ij}$$

$$\hat{KG}_{ijk\ell} = \sum_{m=1}^{R_{ijk\ell}} \frac{H_{ijk\ell}}{R_{ijk\ell}} KG_{ijk\ell m}$$

$$\hat{\bar{KG}}_{ijk} = \sum_{\ell=1}^{\mathcal{L}_{ijk}} \hat{KG}_{ijk\ell} / \mathcal{L}_{ijk}$$

$$\hat{H}_{ijk} = \sum_{\ell=1}^{\mathcal{L}_{ijk}} \frac{C_{ijk}}{\mathcal{L}_{ijk}} H_{ijk\ell}$$

3.3 Estimateurs par la régression

Il existe un potentiel pour réduire la variation entre les producteurs et entre les champs en recueillant de l'information supplémentaire chez tous les producteurs de la zone ou pour tous les champs de chaque producteur échantillonné.

Par exemple, on pourrait mesurer pour chacun des champs d'un producteur échantillonné :

- superficie
- âge du champ
- pourcentage de légumineuses
- nombre de coupes dans l'année
- type de conservation des fourrages

Certains de ces éléments seront effectivement mesurés dans le cadre du projet-pilote et une analyse sera menée afin d'évaluer leur efficacité.

4. Conclusion

Les techniques d'échantillonnage sont fréquemment utilisées à la Régie des assurances agricoles du Québec dans les opérations d'expertise en assurance-récolte.

Pour les programmes d'assurance du système individuel, l'optimisation des tailles d'échantillon à prélever pourrait être envisagée, malgré le caractère imprévisible du niveau de variation entre les unités d'échantillonnage.

Le projet-pilote entrepris dans le programme Grandes cultures selon le système collectif (catégorie foin) permettra de valider le nouveau plan d'échantillonnage proposé et de préciser davantage le nombre d'unités à sélectionner à chacun des degrés d'échantillonnage.

5. Bibliographie

COCHRAN, W.G. (1977). Sampling techniques. Third edition. John Wiley and sons. 428 p.

SESSIONS SUR L'ANALYSE DE DONNÉES

UN SCHÉMA DE VÉRIFICATION BASÉ SUR L'ANALYSE MULTIVARIÉE (1)

France Bilocq, Jean-Marie Berthelot
Statistique Canada (2)

Résumé

Afin de faciliter le développement de la méthodologie pour la fonction générale de collecte, saisie et vérification des données, on fait appel aux techniques d'analyse multivariée. Ces techniques sont utilisées pour concevoir un schéma de vérification pour les variables quantitatives des enquêtes économiques. Des groupements naturels de variables sont identifiés, permettant ainsi d'optimiser l'application des procédures de vérification des données.

1. INTRODUCTION : le problème

La fonction générale de collecte, saisie et vérification est définie comme étant l'ensemble des opérations nécessaires à l'acquisition des données d'enquête, leur validation ainsi que leur conversion sous forme ordinolingue.

La vérification est une étape permettant d'assurer la validité et la cohérence des données fournies par un répondant. Des règles de validité (numérique vs alphabétique, longueur d'un champ etc.) sont appliquées pour s'assurer que l'information saisie correspond à celle fournie par le répondant. Cette information peut ainsi être traitée de façon automatique par la suite. Pour assurer la cohérence entre les variables, on fait appel à des méthodes de détection de données suspectes. Ces méthodes servent à identifier les dossiers dont les variables se comportent différemment du reste de la population étudiée.

Il existe diverses techniques permettant de vérifier si les valeurs des variables d'un dossier sont cohérentes. On peut entre autres, effectuer une vérification historique c'est à dire confronter une variable d'un cycle d'enquête avec la même variable du cycle précédent. On peut aussi confronter deux ou plusieurs variables d'un même cycle c'est à dire, effectuer des vérifications croisées. C'est alors que se pose le problème : pour un questionnaire contenant plusieurs variables financières, il faut, parmi toutes les vérifications croisées possibles, identifier celles qui sont appropriées pour assurer la cohérence d'un dossier.

---

(1) Présenté au colloque sur les applications statistiques du 57e congrès de l'Association Canadienne Française pour l'Avancement des Sciences, Montréal, mai 1989

(2) France Bilocq, Jean-Marie Berthelot, Division des méthodes d'enquêtes-entreprises, Statistique Canada, Ottawa (Ont), K1A 0T6.

Toutefois, ceci n'est pas suffisant pour assurer que le schéma de vérification soit optimal. Pour obtenir un schéma optimal, il faut trouver le patron de vérification permettant de minimiser le nombre de vérifications croisées tout en demeurant efficace et efficient. En d'autres termes, il faut, fournir non seulement un maximum de résultats avec un minimum d'effort mais aussi produire de bons résultats. Le problème étant posé, quelles sont les solutions possibles ?

La première idée qui vient à l'esprit est d'utiliser la matrice des coefficients de corrélation de Pearson. Ce coefficient mesure le degré d'association linéaire existant entre deux variables. En utilisant cette information, on a tenté de regrouper les variables fortement corrélées pour ensuite effectuer, à l'intérieur de chaque groupe, les vérifiations croisées approriées. Toutefois, analyser une matrice dont les dimensions sont grandes n'est pas une tâche facile. De plus il faut un critère permettant de prendre une décision. En effet il est difficile de juger si la corrélation entre deux variables est suffisamment grande pour justifier l'existence d'une relation linéaire. Par exemple, une corrélation de .65 est-elle suffisamment grande pour justifier une vérification croisée entre deux variables ? C'est alors que les techniques d'analyse multivariée viennent à la rescousse.

2. DESCRIPTION DE LA MÉTHODE DE CLASSIFICATION

L'analyse factorielle est une branche de l'analyse des données qui se concentre sur l'étude des relations internes d'un ensemble de variables. La méthode de classification des variables décrite ci-dessous fait appel à cette théorie.

La méthode de classification des variables est un outil statistique qui sert à identifier des groupements naturels de variables permettant d'optimiser l'application des procédures de vérification. Le but de la méthode est de partitionner un ensemble de variables en deux ou plusieurs groupes disjoints. Chacun des groupes formés contient des variables qui sont fortement corrélées les unes avec les autres. Par la suite, un patron de vérification faisant intervenir uniquement les variables à l'intérieur d'un groupe est développé de manière à minimiser le nombre de vérifications croisées. Avant d'entrer dans les détails de la stratégie d'utilisation de la méthode de classification, voici un bref aperçu de cette méthode.

La méthode de classificaton considérée partitionne un ensemble de variables en groupes disjoints en se basant sur leur matrice de corrélation. Les groupes sont formés de façon à maximiser la variance expliquée par la première composante principale de chaque groupe tout en s'assurant que la variance expliquée par la deuxième composante principale n'est pas trop élevée. On contrôle ces deux contraintes à l'aide de paramètres.

Étapes de la méthode de classification :
(Les détails algébriques sont présentés en appendice.)

Au départ, la méthode considère toutes les variables comme faisant partie du même groupe. Par la suite les étapes suivantes sont répétées jusqu'à ce que toutes les conditions soient remplies.

1) La première étape consiste à effectuer l'analyse en composantes principales de la matrice de corrélation de chacun des groupes de variables de façon indépendante.

2) Un groupe donné est subdivisé si la proportion de variance expliquée par sa première composante principale est insuffisante ou si la proportion de variance expliquée par sa deuxième composante principale est trop grande. Autrement, un groupe répondant aux critères n'est pas séparé.

3) Un groupe est séparé en effectuant une variante de la rotation orthoblique proposée par H.H. Harman (1976) [3] et W. Harris et F. Keiser (1964) [4]. Cette technique se nomme ainsi puisqu'on obtient une solution oblique à partir d'une rotation orthogonale et d'une matrice définie positive. Une solution oblique implique que les axes de référence ne sont pas perpendiculaires.

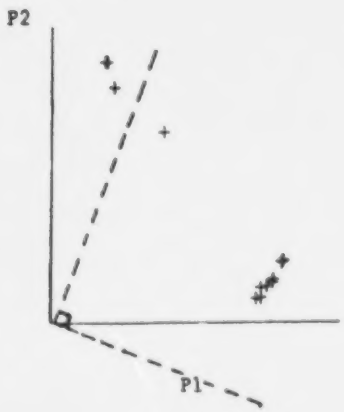
   Un des principes de base de l'analyse factorielle est de toujours être à la recherce d'une solution simple. Par solution simple on entend une solution pour laquelle chaque variable est bien représentée par peu d'axes (de préférence un seul). Or, une telle solution ne se trouve pas toujours automatiquement par l'analyse factorielle d'une matrice de corrélation. C'est pourquoi on fait appel à la rotation.

   L'analyse en composantes principales fournit une des solutions possibles d'axes de référence parmi une infinité. La rotation consiste à faire pivoter ces axes de façon à obtenir une meilleure représentation des variables sur ceux-ci. Cette étape permet de trouver une solution plus claire qui explique plus adéquatement la variabilité d'un groupe de variables.

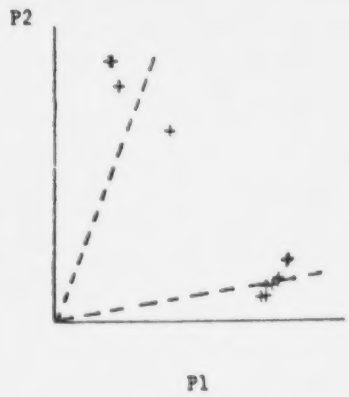
   Pourquoi rechercher une solution oblique plutôt qu'orthogonale ? En imposant un cadre de référence orthogonal on impose aussi la même contrainte aux axes qui le compose. Par exemple, on sait qu'il existe un lien entre les variables concernant les produits manufacturiers et les variables concernant les produits non-manufacturiers (Recensement des manufacturiers). Cependant, ce lien n'est pas assez fort pour justifier le regroupement de ces deux catégories de variables de façon à ce que l'ensemble des variables soient bien expliquées par un seul axe.

Si on impose une rotation dont la solution est orthogonale, on obtient que l'un ou l'autre des groupes est bien expliqué par un des axes au détriment de l'autre. Alors qu'en permettant l'obtention d'une solution oblique on peut simultanément bien expliquer les deux groupes de variables à l'aide d'axes différents. Afin de mieux visualiser la situation transportons-nous dans le graphique des variables représentées dans le plan des deux premières composantes principales.

GRAPHIQUE solution orthogonale



GRAPHIQUE solution OBLIQUE



Note : Les données utilisées pour ces graphiques proviennent du recensement des manufacturiers de 1985.

La solution oblique est plus simple à interpréter. Opter pour une telle solution permet d'obtenir une classification des variables plus détaillée.

Le nombre de groupes obtenu par la méthode de classification se contrôle à l'aide des paramètres spécifiés quant à la proportion de variance expliquée par les deux premières composantes principales. En effet, plus le paramètre contrôlant la première composante principale est grand ou plus celui contrôlant la deuxième composante principale est petit, alors plus le nombre de groupes formés est grand.

Finalement, pour séparer un groupe en deux on regarde les corrélations entre les variables et les deux premiers axes obtenus par la rotation orthoblique. Chaque variable est alors jumelée avec l'axe avec lequel elle est le plus corrélée pour ainsi obtenir deux groupes. Étant donnée que la rotation orthoblique fournit une solution dont les axes ne sont pas orthogonaux, la méthode de classification fournit les coefficients de corrélation entre les groupes.

4) Un processus itératif réassigne (au besoin) les variables dans les groupes afin de maximiser la variance expliquée par la première composante de chacun des groupes.

5) Retour à l'étape 1). Le processus s'arrête lorsque tous les groupes satisfont les critères spécifiés quant à la proportion de variance expliquée par les deux premières composantes principales.

3. STRATÉGIE D'UTILISATION

La fonction générale de collecte, saisie et vérification se divise en deux parties distinctes soit le système de conception et le système de production.

Le système de conception est utilisé par les responsables d'enquête, les méthodologistes et les informaticiens pour définir les besoins d'une enquête donnée en termes de collecte, saisie et vérification. On parle alors de conception et d'impression de questionnaires, d'élaboration d'un plan de vérification, de calcul des paramètres nécessaires aux méthodes de détection etc. Le système de production consiste quant à lui à la mise en marche de la collecte de la saisie et de la vérification des données d'enquête. Cette partie du système est utilisée par les divisions des opérations centrales et régionales. Malgré la distinction faite entre ces deux systèmes il n'en reste pas moins qu'ils sont considérés comme faisant partie d'un produit intégré.

La méthode de classification des variables s'insère donc dans le système de conception. Elle servira d'outil d'analyse en vue d'établir un patron de vérification. Il est important de retenir que la méthode de classification a pour but de regrouper les variables qui sont reliées entre elles pour les besoins de la vérification. Une fois que les groupes sont formés, on peut restreindre les vérifications croisées aux relations qui existent entre les variables d'un groupe tout en assurant un fort degré de cohérence des variables d'un dossier. Des vérifications impliquant des variables de différents groupes seront effectuées uniquement si elles sont nécessaires pour assurer un plus haut degré de cohérence d'un dossier. La procédure de classification fournit les coefficients de corrélation entre les groupes. Cette information pourrait alors servir de guide dans le choix des vérifications inter-groupes.

Lors des études empiriques effectuées sur des données réelles, la procédure VARCLUS de SAS [6] a été utilisée. Cette procédure utilise une technique similaire à celle qui a été présentée dans ce document. Les résultats obtenus lors de ces études sont probants. La méthode de classification extrait facilement le patron de corrélation d'un ensemble de variables. De plus, les groupes de variables obtenus sont représentatifs des caractéristiques spécifiques observées sur les données. Les résultats concordent avec les connaissances des responsables de l'enquête étudiée [2].

La méthode de classification est donc un outil de travail permettant de combiner à la fois théorie statistique et connaissance des données. Un des objectifs de la théorie statistique est de fournir une loi scientifique ou un modèle mathématique pour expliquer le comportement des données. Si l'on possède une connaissance préalable du comportement des variables étudiées, ces notions permettront alors de vérifier la cohérence des groupes formés. De plus, certains liens inconnus entre des variables peuvent être dévoilés grâce à l'application de cette méthode. Cependant, si on ne possède pas cette connaisance ou si l'on veut s'appuyer sur des résultats d'expérience, alors la méthode de classification permet de grouper les variables de façon empirique et objective. Cette méthode permet donc d'enchâsser le processus de vérification des données dans un cadre théorique sans toutefois en oublier le caractère intuitif.

4. BIBLIOGRAPHIE

[1] Berthelot, J.-M. (1989), Approche générale pour la sous-fonction de vérification et de correction des données, Statistique Canada, Division des méthodes d'enquêtes entreprises, Document de travail.

[2] Bilocq. F. (1989), Analyses sur la classification des variables et sur la détection des données suspectes, Statistique Canada, Division des méthode d'enquêtes-entreprises, Document de travail.

[3] Harman, H.H. (1976), MODERN FACTOR ANALYSIS, 3rd ed, University of Chicago Press, Ill. (487 p.)

[4] Harris, W. et Keiser, F. (1964), Oblique factor analystic solutions by orthogonal transformations, PSYCHOMETRIKA, vol 29, no 24, p. 347-362.

[5] Holzinger, K.J. et Harman, H.H. (1942), FACTOR ANALYSIS, University of Chicago Press, Ill. Chicago, Ill., (417 p.).

[6] SAS institue, (1985), SAS USER'S GUIDE : STATISTICS, Version 5, North Carolina.

[7] Thorndike, A.M., (1978), CORRELATION PROCEDURES FOR RESEARCH, Gardner Press, New York.

APPENDICE

Description théorique de la méthode de classification.

Cet appendice contient une description algébrique des étapes de la méthode de classification concernant l'analyse en composantes principales et la rotation orthoblique.

Notation

I : Matrice identité

R : Matrice de corrélation entre les variables

Q : Matrice des vecteurs propres de la matrice R

Λ : Matrice diagonale contenant les valeurs propres de la matrice R

A : Matrice des coefficients (α_{ij}) de l'analyse en composantes principales ("factor loadings")

C_j : $j^{ième}$ composante principale

T : Matrice de rotation orthogonale

S : Matrice de corrélation entre les variables et les axes

Φ : Matrice de corrélation entre les axes

P : Matrice des nouveaux coefficients ("factors loadings") obtenus après une transformation de la matrice A

D : Matrice diagonale définie positive

1) **Analyse en composantes principales**

Une des premières étapes de la méthode de classification est d'effectuer séparémemt, l'analyse en composantes principales (ACP) de la matrice de corrélation de chacun des groupes.

Selon la théorie de l'ACP, on peut trouver A ("factor loadings") de façon à ce que $R = A\,A'$

On sait que $R = Q\,\Lambda\,Q'$ décomposition de la matrice de corrélation en vecteurs et valeurs propres

si on pose $A = Q\,\Lambda^{1/2}$

on a que $A\,A' = Q\,\Lambda^{1/2}\,\Lambda^{1/2}\,Q'$

$= Q\,\Lambda\,Q' = R$

La proportion de variance expliquée par une composante principale (C_j) se calcule de la façon suivante :

Variance expliquée par $C_j = \Sigma_i \alpha^2_{ij} = \lambda_j$

(où λ_j = $j^{ième}$ valeur propre)

La proportion de variance expliquée $P_j = \lambda_j + \Sigma_i \lambda_i$

Si $P_1 < k_1$ ou $P_2 > k_2$ alors le groupe doit être séparé.
(k_1 et $k_2$sont des paramètres déterminés à l'avance).

2) Rotation orthoblique

Par la suite, un groupe est séparé en effectuant une variante de la méthode de rotation orthoblique proposée par Harman [3]. Cette méthode se nomme ainsi puisqu'on obtient une solution oblique par une transformation composée d'une matrice de rotation orthogonale et d'une matrice diagonale définie positive.

Avant d'expliquer la théorie sous-jacente à cette méthode, un bref aperçu des méthodes othogonale et oblique est présenté.

Dans le cas orthogonal, la matrice A (coefficients de l'ACP) est tranformée à l'aide d'une matrice de rotation orthogonale T (où T'T=TT'=I). On a alors AT=B, où B devient la nouvelle solution orthogonale.

Algébriquement, la corrélation entre une variable (z_i) et un axe (F_j) s'exprime de la façon suivante :

$S = A \Phi \quad (s_{ij} = \rho(z_i,F_j))$

$$\rho(z_i,F_j) = a_{i1}\,\rho(F_j,F_1) + a_{i2}\,\rho(F_j,F_2) + a_{i3}\,\rho(F_j,F_3) + \ldots + a_{ij}\,\rho(F_j,F_j) + \ldots + a_{in}\,\rho(F_j,F_n)$$

Les axes étant perpendiculaires, on a que

$$\rho(F_i,F_j) = 1 \text{ si } i-j \;\text{ --> }\; \Phi = I \;\text{ --> }\; S = A \quad s_{ij}=a_{ij}$$
$$= 0 \text{ sinon}$$

Dans le cas oblique, la matrice A est transformée à l'aide d'une matrice de rotation construite de façon à obtenir une solution dont les axes ne sont pas perpendiculaires. On appelle P la matrice des nouveaux coefficients obtenus, pour ainsi les distinguer des anciens coefficients A. Lors d'une rotation oblique les axes sont corrélés entre eux ($\Phi \neq I$). On a donc que :

$S = P\Phi$

Cas orthoblique

La première étape consiste à appliquer la méthode quartimax [3] sur les deux premiers vecteurs propres afin d'obtenir la matrice de rotation orthogonale T ($TT'=T'T=I$). La méthode quartimax maximise une fonction de 4^e puissance des vecteurs propres afin de déterminer l'angle de rotation θ.

La matrice T se définit alors comme suit :

$T = t_{11} = \cos(\theta)$
$t_{12} = \sin(\theta)$
$t_{21} = -\sin(\theta)$
$t_{22} = \cos(\theta)$
$t_{ij} = 1$ pour $i=j$ $\quad i,j = 3,\ldots,v$
$t_{ij} = 0$ pour $i=j$ $\quad i,j = 3,\ldots,v$ (où v = nombre de variables)

Par la suite il faut trouver D^{-1} une matrice diagonale servant à normaliser l'expression $T' \Lambda T$ de façon à obtenir la matrice de corrélation Φ avec des 1 sur la diagonale.

Calcul de D^{-1}

On veut que $\Phi_{ij} = 1$ si $i=j$

on a que $\Phi = D^{-1} T' \Lambda T D^{-1}$

Posons diagonale $(T' \Lambda T) = D^2$

$$\sigma^2 = D^2$$
$$\sigma = D$$
$$1/\sigma = D^{-1}$$

$$[\text{diagonale } (T' \Lambda T)]^{-1/2} = D^{-1}$$

$$\rightarrow \quad D^{-1} T' \Lambda T D^{-1} = \frac{\sigma_{ij}^2}{\sigma_i \sigma_j} \quad ; \quad \text{si } i=j \rightarrow \quad \frac{\sigma_i^2}{\sigma_i \sigma_i} = 1$$

Une fois que T et D on été calculés, la solution oblique complète se définit comme suit :

$P = Q\ T\ D$	(nouveaux coefficients)
$\Phi = D^{-1}\ T'\ \Lambda\ T\ D^{-1}$	(matrice de corrélation entre les axes)
$S = Q\ \Lambda\ T\ D^{-1}$	(matrice de corrélation entre les variables et les axes)

Les s_{ij} sont utilisés pour associer chacune des variables soit au premier axe ou soit au deuxième axe. Les résultats de cette étape donnent deux groupes de variables disjoints. L'ensemble de la procédure sera appliqué de nouveau sur les groupes jusqu'à ce qu'ils respectent tous les critères imposés sur P_1 et P_2.

La procédure "VARCLUS" de SAS utilise une technique similaire pour effectuer la rotation orthoblique.

Diagnostic quantitatif et qualitatif des zones de risques en scieries

Esther Cloutier[1] et Lucie Laflamme[2]

Introduction

La fabrication du bois de sciage constitue un secteur d'activité où les risques d'accidents sont importants. On y retrouve de plus, d'une entreprise à l'autre, une organisation technique de la production qui peut varier considérablement. Le sciage s'effectue à partir de processus tantôt relativement artisanaux, peu automatisés, fortement consommateurs de main-d'oeuvre et de faible capacité de production; tantôt dans des processus hautement automatisés, requérrant peu de main-d'oeuvre et dotés d'une capacité élevée de production; tantôt enfin dans des entreprises modérément consommatrices et de machinerie et de main-d'oeuvre, dont le volume de production est moyennement élevé.

La fabrication du bois d'oeuvre est un processus en flux continu. Aussi, en plus des équipements requis pour sa transformation, la production nécessite l'utilisation de démêleurs et convoyeurs pour le tri et le transport de la matière ligneuse d'une machine à l'autre, notamment à cause de la dimension importante et variable des pièces de bois.

Il a déjà été observé que les accidents du travail ne sont pas aléatoirement distribués à l'intérieur même des usines du sciage, leur fréquence et leurs caractéristiques variant d'une phase à l'autre du processus de transformation (Cloutier, Laflamme, 84; 89).

Dans l'étude présentée ici, l'objectif général est d'apprécier le risque relatif d'accidents (fréquence, gravité, circonstances) dans divers types d'entreprises, selon leur mode d'organisation de la production, en considérant leur consommation combinée de main-d'oeuvre et de machinerie. Nous voulions aussi voir dans quelle mesure les besoins en prévention se ressemblent ou se différencient d'un type de processus à un autre.

Matériel et méthode

Neuf entreprises ont participé à l'étude. Chacune a fait l'objet d'observations sur les lieux de travail. Au cours des visites, des renseignements généraux sur l'oganisation de la production et du travail ont été colligés et des observations de machines et d'équipements ont été réalisées. Ces

---

1 Chercheure à l'Institut de recherche en santé et en sécurité du travail du Québec (IRSST), 505 Blv de Maisonneuve Ouest, Montréal, Québec, H3A 3C2 (Canada)

2 Professionnelle scientifique à l'Institut de recherche en santé et en sécurité du travail du Québec (IRSST) et chercheure invitée à l'Institut suédois de recherche en santé occupationnelle (ARBETSMILJOINSTITUTET)

entreprises ont été classées en trois types de processus, en considérant le nombre de machines et de travailleurs qu'elles employaient. Ce rapport main-d'oeuvre/machinerie a été normalisé pour un volume et une durée de production constants.

Tous les accidents ayant occasionné des pertes de temps d'une journée de travail et plus (778 cas), inclus aux registres des entreprises depuis janvier 1986 jusqu'au moment de la visite, à l'automne 1987 ont été analysés. Plus de 60 variables ont été utilisées pour les décrire. Elles concernent le lieu de l'accident, le moment où il s'est produit, le travailleur accidenté, les circonstances de l'accident, la blessure subie et des recommandations pour la prévention. Ces recommandations ont été obtenues à partir de deux sources: 1) les recommandations inscrites aux formulaires d'enquête d'accidents des entreprises; 2) celles qui ont été formulées ultérieurement par les ingénieurs participant au projet.

Deux indicateurs de fréquence et de gravité d'accidents ont été employés: le taux d'incidence des accidents[3] (T.I.), qui informe sur le nombre moyen d'accidents par travailleur pendant la période étudiée et le taux annuel d'absence par travailleur[4] (T.A.A.) qui indique pour sa part le nombre moyen de jours de travail annuellement perdus par travailleur, pour cause d'accident du travail.

Des scénarios d'accidents ont été obtenus en utilisant deux techniques d'analyse multidimensionnelle: l'analyse factorielle des correspondances (AFC) et la classification ascendante hiérarchique (CAH) (Greenacre, 1984; Lebart, Morineau, Warwick, 1984). Ces scénarios sont des groupes homogènes d'accidents se produisant dans des circonstances semblables et ayant les mêmes caractéristiques. La relation entre scénarios d'accidents et type d'entreprises a été mesurée par un test du Chi-deux.

Résultats

SITUATIONS TYPES D'ACCIDENTS

Un total de 10 classes homogènes d'accidents ont été reconstituées. Elles peuvent être regroupées en trois catégories. Ceux qui se produisent: 1) en cours de production et de déblocage de billes ou planches; 2) lors des travaux d'entretien ou de réparation et 3) les accidents associés à des activités et lieux de travail variables.

Activité de production et de déblocage

La préparation des billes (68 cas): Cette classe regroupe des accidents survenant aux démêleurs à billes, aux tronçonneuses et aux écorceuses, à

---

3 T.I.= Nb. de lésions compensées/Nb. de travailleurs exposés

4 T.A.A.= Nb. de jours perdus pour accident/Nb. de travailleurs exposés

l'entrée de la scierie. Les zones d'intervention critiques sont les passerelles, l'entrée des machines ainsi que les démêleurs et les convoyeurs. Les opérateurs effectuent alors du déblocage de billes, manuellement ou à l'aide de scies mécaniques, de gaffes, de pics ou de tourne-grumes; tirent, poussent ou roulent une pièce de bois. Pendant ces interventions, ils sont frappés par l'objet ou l'outil manipulé. Les sièges de lésions sont divers (bras, mains, jambes, dos).

Les recommandations formulées consistent à porter une attention particulière aux méthodes de travail, à cadenasser les machines lors d'interventions de déblocages et à améliorer le mode de circulation du bois entre les machines.

Le sciage, le classage, l'empilage et le rabotage (116 cas): Cet autre groupe d'accidents surviennent aux refendeuses, démêleurs à planches, classeurs et déchiqueteuses. Ils affectent les opérateurs de ces différentes machines. Ils se produisent à l'entrée ou à la sortie des machines, aux convoyeurs s'y trouvant. Les blessures se produisent lors d'activités de déblocage (bois coincé, accumulé ou mal placé), alors que les travailleurs manipulent une bille (une planche) ou démêlent manuellement du bois. Ils subissent des contusions et des écrasements en se coinçant entre une bille (ou une planche) et une partie de machine (ou une courroie) ou en étant frappés par un morceau de bois.

Pour prévenir ces accidents il est suggéré de modifier l'aménagement des postes de travail et de revoir le mode de circulation de la matière ligneuse entre les machines pour réduire les possibilités de blocage.

L'éboutage (45 cas): Des accidents se produisent aussi à l'entrée des ébouteuses. Les travailleurs concernés, des ébouteurs, trieurs et classificateurs, se blessent pendant la production, en s'étirant le bras, au moment où ils s'apprêtent à tirer ou pousser des planches. Ces gestes entraînent des efforts excessifs ou des coincements des mains ou des doigts entre deux objets (du bois et une partie de machine ou deux parties de machine). Dans plusieurs cas, du bois accumulé ou mal placé a incité le travailleur à poser le geste qui a occasionné la blessure. Lorsque ces accidents surviennent, les machines sont en opération et, dans plus du cinquième des cas, des contraintes d'aménagement du poste de travail ont posé problème dans les circonstances de l'accident.

Pour prévenir de tels accidents, il est recommandé d'améliorer l'aménagement des postes de travail et la circulation de la matière première sur les convoyeurs.

Travaux d'entretien ou de réparation

Dans l'usine (76 cas): Ce groupe d'accidents est constitué de blessures aux membres supérieurs (bras, mains, doigts) survenant pendant l'entretien ou la réparation d'équipements. Les travailleurs impliqués sont des mécaniciens et des électriciens. Les tâches critiques pour ce type de lésions sont soit l'ajustement et le changement des scies (ou couteaux) soit le nettoyage autour des machines. Pendant ces tâches, les travailleurs heurtent une partie de machine; ils se coupent ou subissent une contusion. Dans certains

cas, ils se blessent en utilisant un marteau, un pic à froid, une pelle ou un grattoir. Le manque d'outil d'appoint, l'aménagement des postes de travail et le fait d'avoir à grimper sur les machines pour accomplir ce type de tâches constituent des contraintes importantes qui en compromettent la sécurité.

Pour réduire l'occurrence de ces accidents, il est suggéré de favoriser l'utilisation d'outils ou d'équipements adéquats, pour faciliter l'accomplissement du travail.

Dans les ateliers (66 cas): Des accidents se produisent aussi aux garages et aux ateliers de peinture, de soudure et à la limerie. Lorsque ces accidents surviennent, les mécaniciens et les soudeurs, sont occupés à l'entretien et à la réparation d'équipements, manipulent un outil, soudent, meulent, attachent quelque chose. La plupart du temps, ils se coupent les bras, les mains ou les doigts, après avoir été frappés par l'outil ou l'objet manipulé ou un autre objet. Dans quelque cas, ils reçoivent des particules dans les yeux ce qui provoquent des irritations.

L'usage d'équipement individuel de protection (gant, lunette) est la recommandation la plus fréquente pour la prévention de ces accidents.

L'entretien des lieux de travail (43 cas): Des blessures surviennent enfin au moment où les journaliers effectuent du nettoyage sous les machines, à l'aide de pelles ou d'accessoires divers. Ils se blessent en se coinçant entre deux objets ou parties de machines. Des problèmes d'environnement (poussière, sciure), d'aménagement des lieux et de posture sont soulignés dans la plupart de ces cas.

Pour réduire l'occurrence de ces lésions il est recommandé d'améliorer les fonctions principales des machines et de revoir l'aménagement des postes de travail.

Accidents se produisant à divers endroits de l'usine

Les chutes en cours de déplacements (81 cas): Au moment où surviennent ces accidents les travailleurs circulent dans l'usine, où encore montent ou descendent d'une passerelle, d'un escalier ou d'un marche-pied. Ils s'infligent des foulures aux membres inférieurs (jambes, chevilles, pieds) suite à une chute. Dans de nombreux cas, l'encombrement des lieux, un plancher inégal, la neige, la glace et le fait d'avoir à grimper sur une pièce d'équipement contribuent à occasionner ces chutes.

Pour améliorer la sécurité pendant ces déplacements, il est recommandé d'être prudent; d'effectuer l'entretien préventif de l'équipement et de nettoyer régulièrement les postes et espaces de travail.

Accidents à proximité d'un camion ou d'une chargeuse (53 cas): Ces accidents affectent des opérateurs de chargeuses ou des camionneurs lorsqu'ils montent, descendent de leur véhicule ou lorsqu'ils se tiennent ou s'accrochent à une partie de machine. Ces gestes divers conduisent à des efforts excessifs, à des chutes ou à des heurts. Les principaux sièges de lésions

atteints sont les membres inférieurs. La neige, la glace ou l'obligation d'avoir à grimper sur quelque chose pour accomplir une tâche joue un rôle dans la survenue de ces accidents. Par ailleurs, plusieurs déclarations d'accidents soulèvent des problèmes de posture et de vibration comme facteur contribuant à ces blessures.

Il est recommandé pour réduire leur occurrence, d'effectuer l'entretien régulier des lieux de travail.

Les douleurs au dos (116 cas): Ce groupe est constitué de blessures au dos provoquées par des efforts excessifs en transportant, tirant ou poussant du matériel, au moment où le travailleur manipule un objet ou effectue un déblocage manuel. Les gestes critiques consistent à soulever des planches, un paquet de bois ou de l'équipement. Les travailleurs les plus exposés à ce genre de lésion sont les journaliers et les empileurs.

Pour les prévenir, il est suggéré de fournir de la formation et de l'information sur les méthodes de travail et de rendre accessibles des outils et de l'équipement adéquat facilitant la tâche du travailleur.

Les irritations des yeux (67 cas): Ces accidents sont des cas où les travailleurs ont les yeux irrités par des particules ou de la poussière lorsqu'ils sont à côté (ou sous) d'une machine (déchiqueteuse, déligneuse, refendeuse, etc.). La plupart du temps ils sont en production ou ils effectuent du nettoyage. Des problèmes d'environnement de travail (vent, poussière) ont joué un rôle dans les circonstances de plusieurs de ces accidents.

Pour les prévenir, il est suggéré de porter des lunettes de sécurité; d'améliorer les systèmes d'aspiration et de ventilation et d'effectuer l'entretien des lieux de façon régulière.

CAPACITÉ DE PRODUCTION DES ENTREPRISES, FRÉQUENCE ET GRAVITÉ DES ACCIDENTS

La capacité de production a été estimé à partir du nombre de travailleurs par machine. Ce rapport a été normalisé pour un volume de production fixes. Les entreprises ont ainsi été réparties en trois catégories. Les entreprises du premier type, qui ont une faible capacité, sont celles où le rapport du nombre de travailleurs sur le nombre de machines est élevé et où le nombre de machines est faible. Les entreprises moyennes sont celles où le rapport entre travailleurs et machines est moins important, mais où le nombre de machines est inférieur à 50. Les entreprises de capacité élevée ont aussi un rapport travailleurs/machines modéré mais elles ont un nombre de machines supérieur à 50.

Le tableau 1 indique que l'incidence des accidents du travail de même que le taux annuel d'absence par travailleur sont plus élevés dans les entreprises dont la capacité de production est élevée, puis dans celles où cette capacité est faible. Ces taux sont moins élevés dans les entreprises moyennes.

Tableau 1 : **Incidence et gravité des accidents selon la capacité de production des entreprises**

Capacité de production	T.I.	T.A.A.
Faible	0,40	3,71
Moyenne	0,28	1,64
Élevée	0,73	4,24
Total	**0,46**	**2,95**

La capacité des entreprises a été appréciée à partir du nombre de travailleurs et de machines employés dans chacune d'elles (ce rapport a été normalisé pour un volume de production et une durée de production constants). Les entreprises ont été réparties en 3 catégories: 1) capacité de production faible, 2) capacité moyenne et 3) capacité élevée.

Cela signifie que le risque relatif d'accidents est à son plus bas dans les entreprises moyennes, consommatrices modérées de main-d'oeuvre par machine de même que de machinerie. Il est plus élevé dans les petites entreprises, employant peu de machines mais un nombre élevé de travailleurs pour chaque machine, en moyenne. Il est encore plus important dans les entreprises de forte capacité de production, moins consommatrices de main-d'oeuvre mais possédant beaucoup de machines.

CLASSES D'ACCIDENTS EN FONCTION DE LA CAPACITÉ DE PRODUCTION

Il existe par ailleurs une relation statistiquement significative entre classes d'accidents et types d'entreprises, tel que l'indique le tableau 2. Ce tableau reproduit la distribution des accidents, en pourcentage, dans chaque classe d'accidents et dans chaque type d'entreprise. On y observe que les entreprises de capacité de production élevée enregistrent un pourcentage d'accidents plus important qu'attendu dans les cas où les travailleurs se blessent au dos et aux yeux, au moment de l'entretien et de la réparation, dans les ateliers et à proximité des véhicules. Les accidents se produisent en cours de production et de déblocage n'y sont toutefois pas plus important qu'attendu.

Les entreprises de capacité de production moyenne cumulent un nombre élevé d'accidents dans deux autres classes d'accidents: ceux qui surviennent en cours de déblocage, à l'entrée de l'usine, et ceux qui surviennent pendant le nettoyage sous les machines.

Tableau 2 : Classes d'accidents et capacité de production des entreprises

Classe d'accidents	Capacité de production			Total
	Élevée	Moyenne	Faible	
Production et déblocage				
Arrivée des billes	44,1	**41,2**	14,7	100,0 68
Sciage, classage, empilage et rabotage	41,3	32,8	**25,9**	100,0 116
Éboutage	60,0	22,2	17,8	100,0 45
Entretien et réparation				
Entretien et réparation d'équipements	53,9	22,4	**23,7**	100,0 76
Entretien et réparation dans les ateliers	**71,2**	16,7	12,1	100,0 66
Entretien des lieux	30,2	**51,2**	18,6	100,0 43
Lieux divers				
Chutes en déplacements	54,3	29,6	16,1	100,0 81
Proximité des véhicules	**75,5**	24,5	0,0	100,0 53
Douleurs au dos	**64,7**	20,6	14,7	100,0 116
Irritations des yeux	**64,2**	34,3	1,5	100,0 63
Total %	**55,8**	**28,7**	**15,5**	**100,0**
n	**408**	**210**	**113**	**731**

Chi-Deux= 69,9 ; d. l.= 18 ; P= 10^{-4}

Enfin, les entreprises à faible capacité de production comptent plus d'accidents qu'attendu, en cours de production ou de déblocage, au sciage, au classage, à l'empilage et au rabotage ainsi qu'en cours d'entretien et de réparation des équipements, dans l'usine.

Par ailleurs, deux classes d'accidents sont proportionnellement bien représentées dans chaque type d'entreprises: les chutes en cours de déplacement et les accidents à l'éboutage.

Discussion et conclusion

RISQUES INTRA-ENTREPRISES

Cette étude met d'abord en évidence le fait que les circonstances et fréquences d'accidents peuvent varier d'une phase à une autre d'un processus de production, tout type d'entreprise confondu.

Ils indiquent de plus que l'activité de travail réalisée par un travailleur lors de l'accident de même que le lieu d'exercise de cette activité constituent des facteurs déterminants des lésions subies, qu'il s'agisse de leur fréquence ou de leurs caractéristiques.

RISQUES INTER-ENTREPRISES

L'étude met de plus en évidence le fait que les risques d'accidents ne sont pas aléatoirement distribués, dans un même secteur d'activité, la fréquence, la gravité et les circonstances d'accidents varient significativement selon le mode d'organisation de la production.

Il semble en fait y avoir un seuil, dans le rapport entre le nombre de travailleurs et le nombre de machines, au-delà duquel les risques à la sécurité du travail augmentent. Ils augmenteraient soit parce que beaucoup de main-d'oeuvre est requise pour assurer le bon fonctionnement du processus et de la machinerie, soit parce que cette main-d'oeuvre est insuffisante et que le débit de production est élevé. A chacun de ces deux pôles, les accidents se sont avérés plus fréquents et plus graves.

De plus, les situations d'accidents ne sont pas les mêmes d'un type de processus à un autre. Certains risques s'avérent plus ou moins importants selon le mode de production considérée. Il semble en outre que plus le processus est consommateur de machinerie plutôt que de main-d'oeuvre, plus les accidents se produisent lors d'activités de travail connexes à la production (entretien des lieux et des équipements). Par contre, il faut souligner que certains scénarios d'accidents demeurent aussi fortement représentés qu'attendu, quel que soit le type d'entreprises.

SITUATIONS D'ACCIDENTS

Cette étude identifie de plus un ensemble de situations types d'accidents dont il a été possible d'apprécier la distribution par catégorie d'entreprises et auxquelles des recommandations ont pû être associées. Plusieurs de ces recommandations s'adressent à l'organisation technique de la production,

qu'il s'agisse de la conception des machines ou du design des espaces de travail ou encore de la fourniture et de l'entretien d'outils et équipements adaptés aux tâches effectuées par les travailleurs des scieries.

Plusieurs situations d'accidents, de l'avis des représentants à la sécurité des entreprises de même que des ingénieurs, pourraient être prévenues par l'amélioration fonctionnelle du processus et par son aménagement. En effet, certaines situations d'accidents sont plus immédiatement liées à des catégories particulières d'entreprises. Toutes cependant ont pu être assorties de recommandations pouvant en faciliter la prévention.

Cette étude suggère donc un certain nombre de voies d'analyse des accidents, tant pour en clarifier les circonstances d'occurrence que pour en prévenir la survenue ultérieure. En outre, les recommandations auxquelles elle a conduit pourront s'inscrire dans des programmes de prévention propres à chaque entreprise et être implantées, à plus ou moins brève échéance, en fonction des priorités que chacune d'elle s'établira.

Bibliographie

Cloutier, E., Laflamme, L., 1984. Analyse de 94 accidents du travail survenus en scierie entre le 1er janvier et le 31 octobre 1983, Québec, Institut de recherche en santé et en sécurité du travail du Québec, coll. Notes et rapport scientifiques et techniques, no. 001, 143 p.

Cloutier, E., Laflamme, L., 1989. Organisation de la production et accidents du travail en scierie: Étude comparée d'entreprises et de phases de production, Québec, Institut de recherche en santé et en sécurité du travail du Québec, coll. Notes et rapport scientifiques et techniques, no. 026, 23 p.

Greenacre, M. J., 1984. Theory and application of correspondence analysis, London, Academic Press.

Lebart, L., Morineau, A., Warwick, K. M., 1984. Multivariate descriptive statistical analysis, Wiley, 231p.

REPRESENTATION BIPLOT ROBUSTE

par

Gaétan Daigle

Université Laval

Cité Universitaire

Québec, Canada

G1K 7P4

RESUME: Dans ce travail, on présente une nouvelle version du biplot basée sur les M-estimateurs multivariés. On considère premièrement une matrice $\mathbf{X}$ de dimension $n \times p$ comme un échantillon de taille n provenant d'une distribution elliptique définie dans $\mathbb{R}^p$. On calcul ensuite les M-estimateurs robustes des paramètres de localisation et d'échelle de cette distribution. On construit le biplot en assignant à chacune des lignes de la matrice $\mathbf{X}$ un poids déterminé au préalable lors de l'estimation robuste des paramètres. Sur le biplot, la longueur du vecteur associé à une variable sera proportionnelle à la variance robuste de cette variable alors que le cosinus de l'angle entre deux variables sera approximativement égal à leur corrélation robuse.

I. Introduction

Dans une analyse en composantes principales (ACP), on représente souvent les individus en les projetant sur le plan formé par les deux premiers axes factoriels. Cette représentation contient la majeure partie de la variabilité dans les données et nous permet de faire sortir les traits dominants de l'échantillon. On appelera ce graphique, le *graphique des individus*. Un autre graphique intéressant en ACP est celui des variables. Il nous permet d'exhiber la structure de la matrice de variances-covariances échantillonale; la norme euclidienne du vecteur associé à une variable est alors proportionnelle à l'écart-type de cette variable tandis que le cosinus de l'angle entre deux variables est approximativement égal à leur corrélation. Ces graphiques des variables originent des travaux de Benzécri (voir Lebart, Morineau et Warwick, 1984). Cependant, Gabriel (1971) a été le premier à montrer que ces graphiques des variables étaient une composante du biplot.

L'ACP standard n'est pas robuste à la présence d'observations extrêmes dans l'échantillon. En effet, il est bien connu que de telles observations peuvent influencer indûment le choix des premières composantes principales, faisant ainsi du graphique des individus le témoin visuel de l'opposition entre les observations extrêmes et le reste des données. Pour représenter la structure de la majeure partie des observations sur le graphique des individus, on doit faire une ACP robuste, laquelle est basée sur les

M-estimateurs multivariés des paramètres de localisation et d'échelle de l'échantillon des individus. Une version robuste du graphique des individus est obtenu en projetant chacun des points de l'échantillon dans le plan formé par les deux premiers vecteurs propres du M-estimateur du paramètre d'échelle de l'échantillon.

Les graphiques des variables sont aussi sensibles à la présence d'observations extrêmes. Il est bien connu que, même en petit nombre, ces observations peuvent avoir un effet considérable sur la structure de la matrice de variances-covariances. En effet, ces observations peuvent quelques fois suggérer une relation entre les variables qui n'est pas représentative de la structure principale de l'ensemble de données. Ce travail a pour objet de suggérer une alternative robuste au graphique des variables et au biplot qui lui est associé.

Notre biplot robuste est basé sur les M-estimateurs des paramétres de localisation et d'échelle de l'échantillon des individus. Le graphique robuste des individus, présenté ci-haut, peut être vu comme une composante du nouveau biplot. La contrepartie de ce biplot est le graphique robuste des variables. Sur ce dernier graphique, la longueur d'un vecteur associé à une variable est reliée à l'écart-type robuste de cette variable tandis que l'angle entre deux vecteurs dépend de la corrélation robuste entre les variables correspondantes.

Les M-estimateurs des paramètres de localisation et d'échelle sont revus brièvement dans la section II. Le biplot robuste est introduit à la section III. La présentation suit celle de Gabriel (1971) qui a montré que le biplot était rendu unique lorsqu'une métrique définissant les distances entre les individus, ou entre les variables était sélectionnée. Le biplot robuste est obtenu en choisissant une métrique bien particulière. Un exemple est présenté dans la section IV.

II. M-estimateurs multivariés des paramètres de localisation et d'échelle

Les M-estimateurs multivariés ont été introduit par Maronna (1976). Ils peuvent être vus respectivement comme une moyenne pondérée des observations et une matrice de variances-covariances pondérée. Le poids de chacune des observations dépend de leur distance par rapport au M-estimateur de localisation. Soit $\{\mathbf{y}_i; i = 1, \ldots, n\}$ un échantillon de taille n d'une distribution multivariée définie dans R^p. Les M-estimateurs, $\hat{\boldsymbol{\mu}}$ et $\hat{\boldsymbol{\Sigma}}$, des paramètres de localisation $\boldsymbol{\mu}$ et d'échelle $\boldsymbol{\Sigma}$ sont définis par les solutions du système d'équations non linéaires suivant:

$$\begin{aligned} \boldsymbol{\mu} &= \frac{\sum_{i=1}^{n} \sqrt{\omega_i}\mathbf{y}_i}{\sum_{i=1}^{n} \sqrt{\omega_i}} \\ \boldsymbol{\Sigma} &= \frac{\sum_{i=1}^{n} \omega_i(\mathbf{y}_i - \boldsymbol{\mu})(\mathbf{y}_i - \boldsymbol{\mu})'}{(n-1)CT}, \end{aligned} \tag{2.1}$$

où $\omega_i = w\left[(\mathbf{y}_i - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{y}_i - \boldsymbol{\mu})\right]$ est une fonction, en général décroissante, de $(\mathbf{y}_i - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{y}_i - \boldsymbol{\mu})$, la distance de Mahalanobis entre $\mathbf{y}_i$ et $\boldsymbol{\mu}$ et

$$CT = E\left\{Yw(Y)\right\}/p. \tag{2.2}$$

Y étant une variable aléatoire chi-deux à p degrés de liberté. La constante CT a été choisie de façon à rendre les M-estimateurs convergeant dans le modèle paramétrique normal (voir Maronna, 1976). Pour une revue des propriétés théoriques des M-estimateurs multivariés, voir Hampel, Ronchetti, Rousseeuw, et Stahel (1986).

Dans le calcul des M-estimateurs, nous avons choisis la fonction du poids de Huber définie par $w(t) = \min(c, t)$ avec $c = p - .5\sqrt{p}$. Nous avons choisis une petite constante pour que nous estimateurs puissent avoir de bonnes propriétés de robustesse. Lorsque $p = 1$, ces estimateurs se réduisent au *Proposal 2* de Huber avec paramètre $\sqrt{.5} \approx .7$ (Huber, 1981). Il est important de noter que l'unicité des solutions du système d'équations défini en (2.1) n'a pas encore été établie sous des conditions de régularité acceptables (Huber, 1981 et Tyler, 1983). Pour calculer nos estimations, nous avons utilisé l'algorithme suggéré dans Huber (1981), p.238. Cet algorithme convergeait vers une solution acceptable pour tous les ensembles de données que nous avons analysés.

Pour construire le biplot biplot robuste, nous avons besoin de la matrice diagonale $\mathbf{W}_{n\times n}$ qui renferment les poids normalisés $\omega_i = w\left[(\mathbf{y}_i - \hat{\boldsymbol{\mu}})'\hat{\boldsymbol{\Sigma}}^{-1}(\mathbf{y}_i - \hat{\boldsymbol{\mu}})\right]/((n-1)CT)$ que nous avons utilisés dans le calcul des M-estimateurs. Si les variables ne sont pas mesurés dans les mêmes unités, elles peuvent être standardisées avec les éléments de la matrice $\mathbf{V} = \text{diag}\left\{1/\hat{\boldsymbol{\Sigma}}_{j,j}\right\}$. Pour avoir des résultats intéressants, nous allons travailler sur la matrice $\mathbf{Y}_{n\times p}$ des points centrés. Ainsi, la $i^{\text{ième}}$ ligne de $\mathbf{Y}$ est égal à $(\mathbf{y}_i - \hat{\boldsymbol{\mu}})$.

III. Biplot robuste

La façon dont nous allons développer le biplot suit la construction de Gabriel (1971). Le biplot est une représentation d'une matrice dans un espace vectoriel de dimension 2. Lorsque $p = 2$, toute la matrice est utilisée pour construire le biplot. Lorsque $p > 2$, le biplot sera construit en utilisant une matrice $\mathbf{X}_{n\times p}$ de rang 2 qui approxime bien $\mathbf{Y}$ (lorsque $p = 2, \mathbf{X} = \mathbf{Y}$). La détermination de $\mathbf{X}$ se fera plus tard dans la section.

Le biplot est basé sur la factorisation de $\mathbf{X}$ sous la forme $\mathbf{AB}'$, où $\mathbf{A}$ et $\mathbf{B}$ sont des matrices $n \times 2$ et $p \times 2$ respectivement. Les lignes de $\mathbf{A}$ et de $\mathbf{B}$ correspondent respectivement aux coordonnées des individus et des variables sur le biplot. Toutefois, cette factorisation n'est pas unique. Elle peut être rendu unique, à une transformation

orthogonale près, en imposant la contrainte que les distances euclidiennes entre les lignes de $\mathbf{A}$ représentent les distances entre les individus dans une métrique quelconque $\mathbf{M}$, i.e.

$$\mathbf{AA}' = \mathbf{XMX}' \tag{3.1}$$

(qui est équivalent à $\mathbf{B}'\mathbf{MB} = \mathbf{I}_2$, où $\mathbf{I}_2$ est la matrice identité 2×2); ceci nous donne le biplot des individus. On obtient le biplot des variables en imposant la contrainte que les distances euclidiennes entre les lignes de $\mathbf{B}$ représentent les distances entre les variables dans une métrique quelconque $\mathbf{N}$, i.e.

$$\mathbf{BB}' = \mathbf{X}'\mathbf{NX} \tag{3.2}$$

(ou $\mathbf{A}'\mathbf{NA} = \mathbf{I}_2$). Lorsqu'une métrique est choisie pour représenter soit les individus ou les variables, une autre est alors induite automatiquement pour l'autre composante de biplot, soit les variables ou les individus. Voir Gabriel (1971) pour les détails.

Pour le biplot des individus, qu'il soit robuste ou non, la métrique la plus souvent utilisée est $\mathbf{M} = \mathbf{I}_p$. Lorsque les variables sont mesurées sur des échelles différentes, il peut être important de les standardisées avant de construire la représentation. Dans l'approche robuste, ceci se fait en considérant le biplot de la matrice $\mathbf{XV}^{1/2}$, où $\mathbf{V}$ est la matrice diagonale dont les éléments correspondent aux inverses des variances robustes (voir section II).

Le biplot des variables classique est obtenu en posant $\mathbf{N} = \mathbf{I}_n/(n-1)$. Son équivalent robuste s'obtient en posant $\mathbf{N} = \mathbf{W}$, où $\mathbf{W}$ est la matrice diagonale contenant les poids des observations que nous avons définie dans la section II. Avec ce choix de métrique, si $\mathbf{X}$ est une bonne approximation de $\mathbf{Y}$, nous avons alors, de par (3.2), que

$$\mathbf{BB}' = \mathbf{X}'\mathbf{WX} \approx \mathbf{Y}'\mathbf{WY} = \hat{\mathbf{\Sigma}}. \tag{3.3}$$

Ainsi, la longueur d'un vecteur représentant une variable sur le biplot est relié à la variance robuste de cette variable, alors que la position relative d'un vecteur par rapport à un autre est reliée à la corrélation robuste entre les variables qu'ils représentent. Considérons maintenant la métrique induite pour les individus dans le biplot des variables. Dans le cas classique, il s'agit de la métrique de Mahalanobis (Gabriel, 1971 et Joliffe, 1986, p.77). Ainsi, la distance d'un individu par rapport à l'origine est proportionnelle à $(\mathbf{y}_i - \bar{\mathbf{y}})'S^{-1}(\mathbf{y}_i - \bar{\mathbf{y}})$ où $\mathbf{S}$ est la matrice de variances-covariances échantillonale habituelle. Dans le cas robuste, on obtient, en suivant le raisonnement de Gabriel, la métrique de Mahalanobis robuste. Les individus sont donc éloignés approximativement de $(\mathbf{y}_i - \hat{\boldsymbol{\mu}})'\hat{\mathbf{\Sigma}}^{-1}(\mathbf{y}_i - \hat{\boldsymbol{\mu}})$ unités de l'origine. Campbell (1982) a étudié ces deux distances; il a montré que la distance robuste performait mieux dans la détection des observations extrêmes.

Retournons maintenant au problème de touver la matrice $\mathbf{X}$ qui approxime le mieux $\mathbf{Y}$. Dans le cas standard, $\mathbf{X}$ est habituellement définie comme étant la matrice de rang 2 qui minimise

$$\sum_{i=1}^{n}\sum_{j=1}^{p}(X_{i,j}-Y_{i,j})^2. \tag{3.4}$$

Gabriel (1978) a donné la solution à ce problème de minimisation en décomposant la matrice $\mathbf{Y}$ en valeurs singulières, i.e.

$$\mathbf{Y}=\sum_{i=1}^{n}\lambda_i\mathbf{u}_i\mathbf{v}_i' \tag{3.5}$$

où

1) $\lambda_i \geq \ldots \geq \lambda_r$
2) r est le rang de la matrice $\mathbf{Y}$
3) $\mathbf{u}_i$ et $\mathbf{v}_i$ sont des vecteurs unitaires dans la métrique euclidienne.

Il a montré que la matrice constituée des deux premiers termes de cette décomposition était la solution cherchée, à savoir $\mathbf{X}=\sum_{i=1}^{2}\lambda_i\mathbf{u}_i\mathbf{v}_i'$. Dans le cas robuste, on définit $\mathbf{X}$ par la matrice de rang 2 qui minimise la somme pondérée suivante:

$$\sum_{i=1}^{n}\sum_{j=1}^{p}\omega_i\phi_j(X_{i,j}-Y_{i,j})^2,$$

où $\{\omega_i; i=1,\ldots,n\}$ et $\{\phi_j; j=1,\ldots,p\}$ sont les poids associés aux individus et aux variables respectivement. Ces poids correspondent en fait aux éléments des matrices diagonales $\mathbf{W}$ et $\mathbf{V}$ définies dans la section II. En se servant cette fois de la décomposition en valeurs singulières généralisés (Greenacre, 1984, annexe B), i.e.

$$\mathbf{Y}=\sum_{i=1}^{r}\beta_i\mathbf{p}_i\mathbf{q}_i'$$

où

1) $\beta_1 \geq \ldots \geq \beta_r$
2) r est le rang de la matrice $\mathbf{Y}$
3) $\mathbf{p}_i$ et $\mathbf{q}_j$ sont des vecteurs unitaires dans la norme induite par les matrices $\mathbf{W}$ et $\mathbf{V}$ respectivement.

on montre, en suivant le raisonnement de Gabriel, que $\mathbf{X}=\sum_{i=1}^{2}\beta_i\mathbf{p}_i\mathbf{q}_i'$.

IV. Exemples d'applications portant sur les données de World Statistics in Brief

Le profil socio-économique de 70 pays sera analysé dans cette section. Pour chacun de ces pays, on dispose des variables suivantes:

1) la population totale mesurée en millions d'habitants (POP)
2) le nombre de mortalités infantiles par 1000 naisssances (MOT)
3) le nombre de véhicules motorisés par 100 habitants (MOT)
4) le nombre de médecins par 1000 habitants (DOC)
5) la fraction du total des transactions consacrée à l'exportation (EXP)
6) la densité de la population par km^2 (DEN)
7) le produit national brut par habitant mesuré en devise américaine (PNB).

Les corrélations entre ces variables sont reportées au tableau I. On peut voir que les corrélations robustes sont généralement plus grandes en valeurs absolue que les corrélations classiques. Nous avons représentés au mieux ces corrélations sur les biplots classique et robuste des variables (voir figure 1 et 2).

Tableau 1 *Matrices de corrélations classiques et robustes**

	POP	MOR	MOT	DOC	PNB	DEN	EXP
POP	1	.06 (.13)	-.10 (-.00)	.01 (-.05)	-.09 (-.04)	.03 (.15)	.03 (.13)
MOR		1	-.68 (-.73)	-.75 (-.80)	-.54 (-.60)	-.14 (-.19)	-.21 (-.30)
MOT			1	.63 (.76)	.86 (.89)	.02 (.08)	.21 (.28)
DOC				1	.48 (.59)	.02 (.16)	.18 (.22)
PNB					1	-.02 (.06)	.19 (.22)
DEN						1	-.19 (-.09)
EXP							1

* *Les chiffres entre parenthèses sont les corrélations robustes de Huber*

Figure 1: Biplot classique des variables

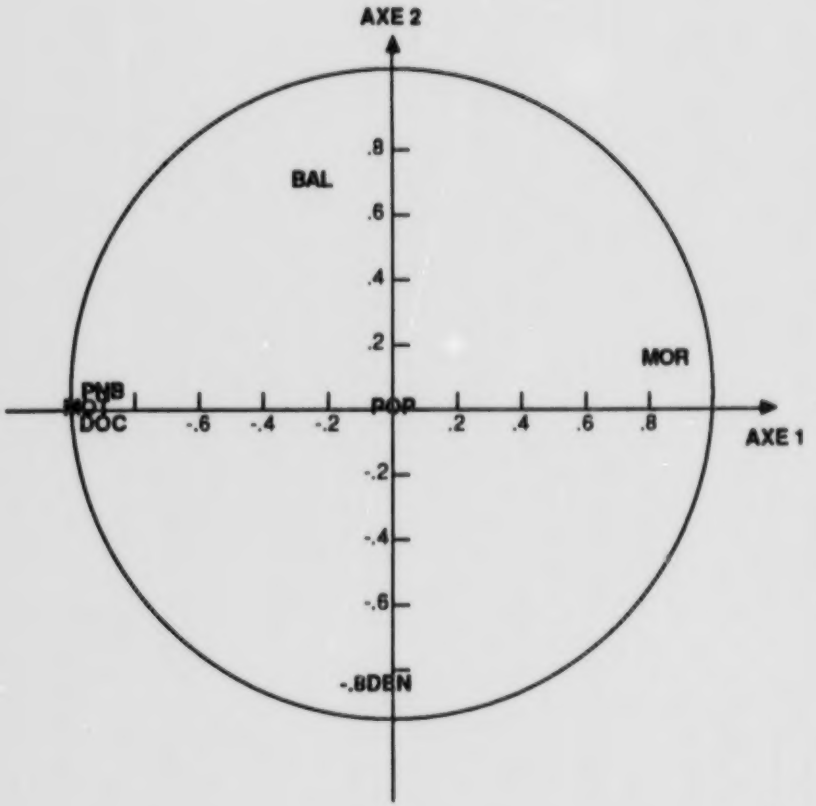
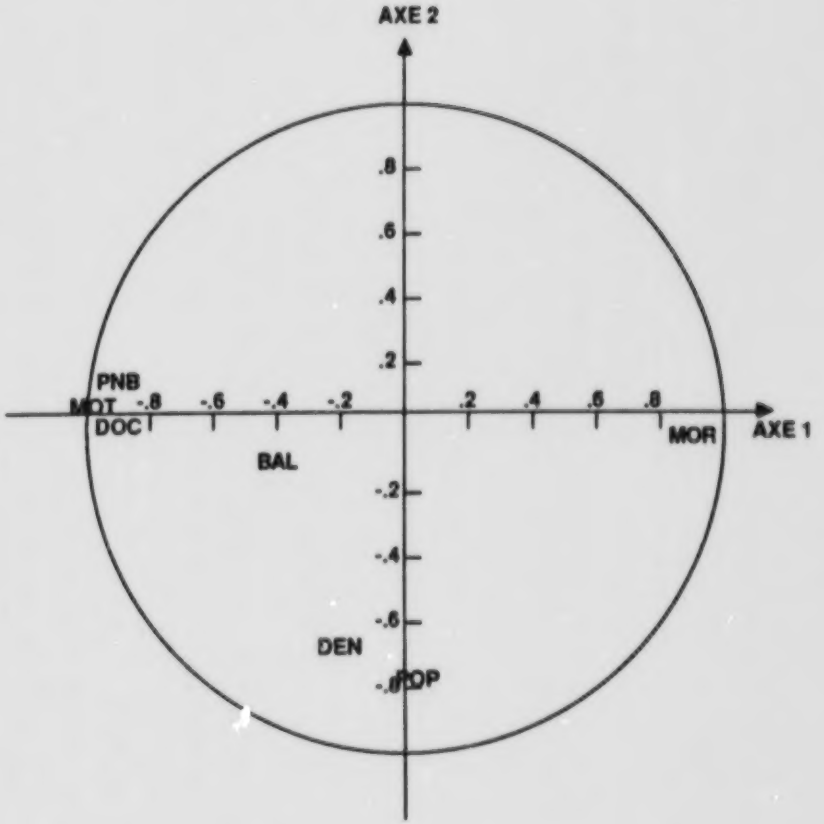


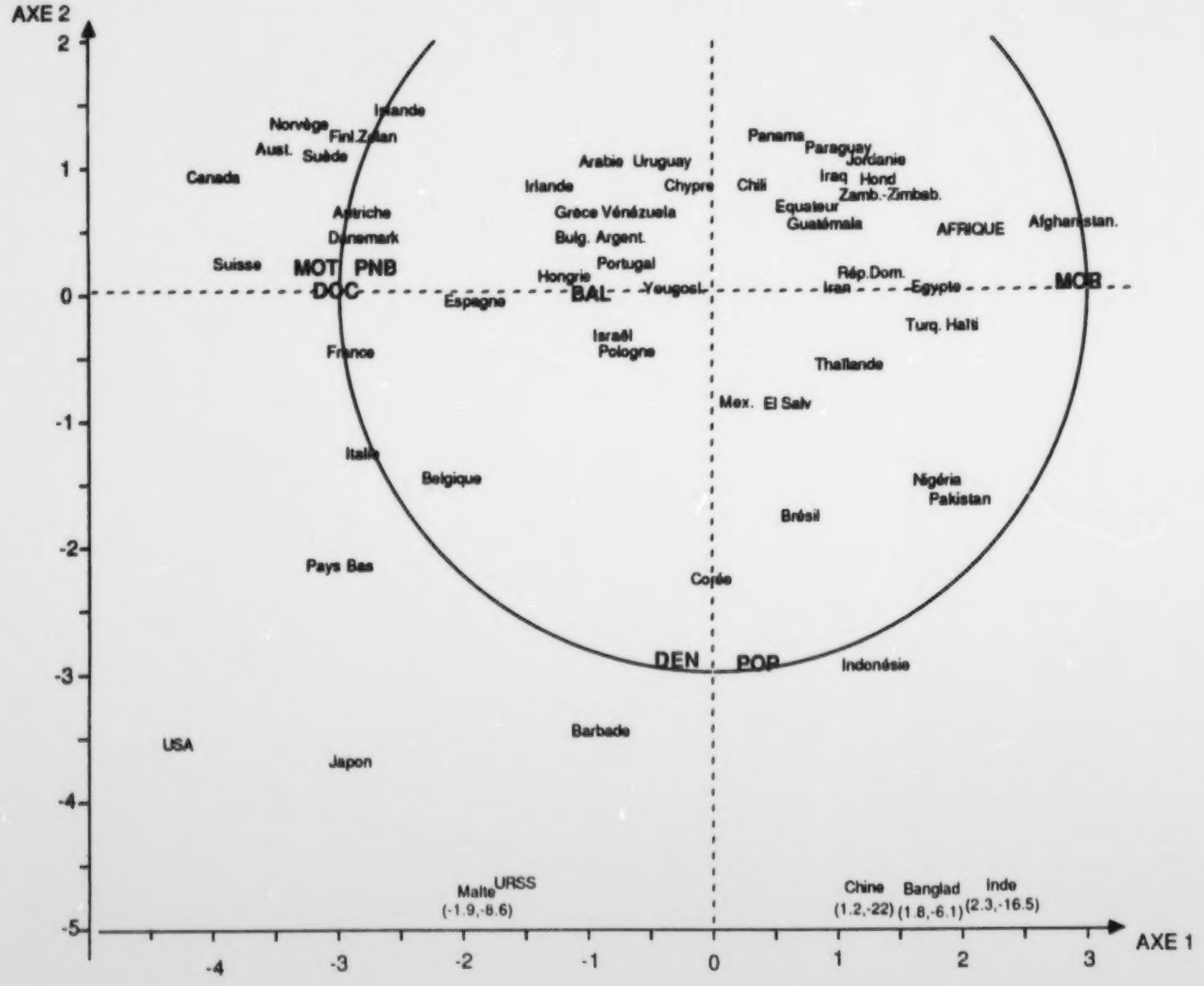
Figure 2: Biplot robuste des variables



Ces deux figures s'entendent en ce qui a trait à l'information apportée par le premier axe factoriel. Cet axe oppose les variables MOT, PNB et DOC à MOR, lesquelles variables expriment en quelque sorte la richesse du pays. Ainsi donc, le premier axe factoriel oppose les pays riches aux pays pauvres. C'est dans la direction du deuxième axe factoriel que le biplot classique diffère du biplot robuste. En effet, le biplot classique laisse voir une relation négative ente EXP et DEN alors qu'on s'attendait à une abscence de relation entre ces variables. La raison en est bien simple; les pays de Malte, du Bangladesh, et de la Barbade sont les seuls à admettre une forte densité de population par km^2 avec un volume de transaction consacrée à l'exportation relativement faible. Ces pays ont considérablement influencé le biplot classique. En contre-partie, on peut voir que le biplot robuste des variables a résisté à l'influence de ces pays. Ce biplot laisse parler la structure principale des observations en montrant une abscence de relation entre DEN et EXP.

Voyons maintenant si on peut regrouper certains pays en classes socio-économiques. Pour ce faire, nous allons utiliser l'information disponible sur le biplot robuste des individus représenté à la figure 3. Les pays qui sont bien représentés sur l'axe factoriel 1 sont soit très riches, soit très pauvres selon qu'ils sont situés du côté négatif ou positif de l'axe. On caractérise un pays riche par un produit national brut assez élevé, un grand nombre de véhicules motorisés et de médecins par habitant, et finalement un faible taux de mortalité infantile. Ainsi, on peut dire que les pays de l'Amérique du Nord, de l'Europe du Nord et de l'Ouest sont généralement riches alors que les pays comme Haïti, l'Afghanistan, la Chine, et l'Inde sont relativement pauvres. Des conclusions d'ordre démographiques peuvent se faire à partir du deuxième axe factoriel. En effet, les pays qui sont bien représentés sur cet axe sont soit très peuplés ou très dense, soit peu peuplés ou peu dense selon qu'ils sont situés sur le côté négatif ou positif de l'axe. Ainsi, les pays comme la Barbade, le Japon, les Etats-Unis, la Chine et l'Inde sont des pays dont la population est relativement importante. Par opposition, les pays comme l'Uruguay, le Paraguay, le Panama, la Jordanie et l'Islande ont soit une faible population, soit soit une grande population mais très éparpillée.

Figure 3: Representation robuste des individus



REFERENCES

[1] Campbell, *Robust Procedures in Multivariate Analysis: Robust Covariance Estimation*, Appl. Statist. **29** (1980), 231- 237.

[2] Gabriel, K.R., *Least Squares Approximations of Matrices by Additive and Multiplicative Models*, J. R. Statis. Soc. B **40** no 2 (1978), 182–196.

[3] Gabriel, K.R., *The Biplot Graphic Display of Matrices with Application to Principal Component Analysis*, Biometrika **58** no 3 (1971), 453–467.

[4] Greenacre, M.J., "Theory and Applications of Correspondence Analysis," Academic Press, London, 1984.

[5] Hampel, F.R., Rousseeuw, P.J., Ronchetti, E.M., Stahel, W.A., "Robust statistics: The Approach Based on Influence Functions," Wiley, New York, 1986.

[6] Huber, P.J., "Robust Statistics," Wiley, New York, 1981.

[7] Joliffe, I.T., "Principal Component Analysis," Spinger Verlag, New York, 1984.

[8] Lebart, L., Morineau, A., Warwick, K.M., "Multivariate Descriptive Statistical Analysis," Wiley, New York, 1984.

[9] Maronna, R.A.,, *Robust M-estimators of Multivariate Location and Scatter*, the Annals of Statistics **4** no 1 (1976), 51–67.

[10] Tyler, D.E., *Robustness and Efficiency Properties of Scatter Matrices*, Biometrika **70** (1983), 411–420.

[11] "World Statistics in Brief, *number 10*," United Nations, New York, 1986.

MESURES DE SIMILITUDE DANS L'ANALYSE DE REGROUPEMENT D'HOPITAUX*

par

Sandra DUFOUR

Professionnelle de recherche

et

Gilles D'AVIGNON

Professeur de systèmes de décision

Rapport de recherche RR-89/8

(juin 1989)

Groupe de recherche sur l'aide à la décision et à l'évaluation

Faculté des sciences de l'administration

Université Laval, Qué. G1K 7P4

---

* Ce texte constitue la base de la présentation de Sandra DUFOUR au colloque "Méthodes et domaines d'application de la statistique 1989" dans le cadre du 57e congrès de L'ACFAS, UQAM, Montréal, 15-19 mai 1989.

SOMMAIRE

INTRODUCTION

PROFILS D'HÔPITAUX

MESURES DE SIMILITUDE

I. Distance euclidienne

II. Distance du khi-carré

III. Distance informationnelle

DISCUSSION

RÉFÉRENCES

SOMMAIRE

Dans la gestion des soins de santé, une préoccupation majeure est l'administration d'un réseau d'hôpitaux publics. Afin d'établir l'efficience des productions hospitalières, il est habituel de former des groupes d'hôpitaux homogènes. La formation d'ensembles homogènes d'hôpitaux peut s'effectuer à l'aide d'une analyse de regroupement. Cette analyse statistique utilise, dans une première phase, une mesure de similitude entre chaque paire d'hôpitaux.

Ce document a pour objectif d'analyser trois mesures de similitude afin d'en choisir une qui soit la plus appropriée au problème qui nous intéresse.

INTRODUCTION

Dans des établissements publics comme les hôpitaux, il est difficile de mesurer l'efficience de leur production puisqu'on ne peut se baser sur les lois du profit comme c'est le cas dans le domaine des entreprises privées.

Mais certains indicateurs ou mesures de performance ont tout de même été construits afin d'évaluer la gestion des différents centres d'activités des hôpitaux. Le coût moyen par catégorie de diagnostic et le séjour moyen sont deux indicateurs couramment utilisés. Toutefois, il n'existe pas de standards permettant d'apprécier la valeur de ces indicateurs.

Le gouvernement, qui finance ces hôpitaux, a donc recours à des comparaisons afin de mieux apprécier leur performance et ainsi déterminer ceux qui réalisent de bonnes performances, et exiger des correctifs pour les centres moins performants.

Au Canada, plus particulièrement au Québec, les comparaisons interhospitalières ont très souvent été fondées sur des classifications arbitraires, telles des variables de ressources. En se basant sur des variables de production, on pourrait comparer les hôpitaux d'après ce qu'ils font et non d'après ce qu'ils sont ou ce qu'ils possèdent comme ressources.

La production hospitalière peut être représentée en partie par le profil des soins prodigués (D'Avignon, Dionne, Dufour [1989]). Ces profils servent d'information de base dans la caractérisation des hôpitaux. Le but d'une analyse de regroupement est d'utiliser ces profils dans la classification des hôpitaux dans des groupes les plus homogènes possibles. Ainsi, les hôpitaux d'un même groupe sont semblables par rapport à leur profil, alors que deux hôpitaux situés dans deux groupes distincts sont relativements différents.

PROFILS D'HÔPITAUX

La détermination du profil d'un hôpital, basé sur des variables de production, peut se faire de différentes façons. A partir de caractéristiques dimensionnelles des patients d'un hôpital, on peut comptabiliser le nombre de patients ayant des caractéristiques spécifiques ou modalités données de patients. Parmi les caractéristiques dimensionnelles que les banques de données sur la production hospitalière nous permettent de considérer (Bourdages [1985]), citons les suivantes:

a) Séjour: nombre de jours durant lesquels le patient a séjourné dans l'hôpital (valeur de 1 à 9999 jours). Cette variable a été codée de façon à prendre une valeur dans l'une ou l'autre des dix classes suivantes: 1 jour, 2 jours, 3 à 4 jours, 5 à 6 jours, 7 à 8 jours, 9 à 10 jours, 11 à 13 jours, 14 à 17 jours, 18 à 89 jours, au moins 90 jours. Ce codage du séjour permet la formation de 10 classes comptant chacune à peu près le même nombre de patients.

b) Age: ici 5 classes d'âge sont utilisées, à savoir: moins d'un an, de 1 à 14 ans, de 15 à 64 ans, de 65 à 74 ans, et finalement 75 ans ou davantage.

c)Traitements: actes diagnostiques, thérapeutiques et chirurgicaux posés au cours de l'hospitalisation. On distingue 19 types de traitements selon la codification canadienne.

d) Diagnostic: il s'agit du diagnostic déterminé par le médecin et considéré par celui-ci comme étant la cause principale de l'hospitalisation. Il y a 20 grandes catégories de diagnostics selon le CIM-9 (Codification Internationale de la Médecine).

Ainsi, pour un hôpital donné, si l'on choisit l'âge et le séjour comme étant deux caractéristiques dimensionnelles, on pourrait dénombrer 42 patients sur les 200 hospitalisations de l'année appartenant à la modalité âge inférieur à 15 et séjour égal à 2.

Il est aussi possible de proposer d'autres caractéristiques dimensionnelles.

Comme on désire grouper les hôpitaux sur la base des caractéristiques des patients qui y ont été hospitalisés, on choisit un certain nombre de ces caractéristiques dimensionnelles des bénéficiaires. Le produit cartésien de ces caractéristiques dimensionnelles nous donne ainsi un profil de modalités. Si l'on considère trois caractéristiques dimensionnelles ayant respectivement 5, 8 et 6 classes, alors on aura un profil comptant 240 modalités pour représenter les hospitalisations des patients dans les hôpitaux. Voulant conserver ou éliminer l'effet de taille d'un hôpital, le profil peut être respectivement donné sous forme de fréquences brutes ou de fréquences relatives par hôpital (Dufour [1988]). Les données de base du problème de regroupement d'hôpitaux peuvent donc être présentées au moyen du tableau 1 qui suit.

Tableau 1. Profils de fréquences relatives ou brutes

	Modalités: M_1	M_2	...	M_k	...	M_m
Hôpitaux:						
H_1	p_{11}	p_{12}	...	p_{1k}	...	p_{1m}
H_2	p_{21}	p_{22}	...	p_{2k}	...	p_{2m}
.						
.						
.						
H_i	p_{i1}	p_{i2}	...	p_{ik}	...	p_{im}
.						
.						
.						
H_j	p_{j1}	p_{j2}	...	p_{jk}	...	p_{jm}
.						
.						
.						
H_n	p_{n1}	p_{n2}	...	p_{nk}	...	p_{nm}

Notation:

m- nombre de modalités dans le profil

M_k- k-ième modalité

n- nombre d'hôpitaux

H_i- i-ième hôpital

p_{ik}- nombre ou proportion des patients de H_i appartenant à la modalité M_k selon que le profil est donné respectivement par des fréquences brutes ou des fréquences relatives.

Par exemple, si l'on prend en considération 2 caractéristiques dimensionnelles, soient le séjour (les 10 classes: J_1 à J_{10}) et le diagnostic (les 19 regroupements majeurs de CIM-9: D_1 à D_{19}), chaque profil est défini par les combinaisons possibles de ces 2 caractéristiques dimensionnelles. Par exemple, le vecteur de fréquences relatives:

	D_1XJ_1	D_1XJ_2	D_1XJ_3	----------	$D_{19}XJ_{10}$
H_1	0,75	0,25	0		0

traduirait le profil de l'hôpital H_1 dans lequel 75% des hospitalisations ont un diagnostic D_1 et un séjour J_1 et dans lequel aussi 25% des hospitalisations ont un diagnostic D_1 et un séjour J_2.

Dans la problématique du regroupement d'hôpitaux, nous comparons en fait des profils de soins, lesquels sont caractérisés au moyen d'un vecteur de fréquences. Une grande similitude entre 2 profils devra signifier que les patients des 2 centres hospitaliers se répartissent à peu près de la même façon selon les différentes modalités considérées.

MESURES DE SIMILITUDE

Il s'agit de choisir un indice de proximité qui traduira les ressemblances ou les dissemblances entre hôpitaux, i.e. un nombre unique pour résumer la similitude entre les profils caractérisant deux hôpitaux (Dionne [1984]). Il y a plusieurs mesures de similitude selon le type de données: nominales, ordinales, d'intervalles ou de rapports.

Notre tableau de données est constitué des profils des hôpitaux; il a donc la forme d'une matrice de fréquences relatives ou d'une matrice de contingence si elle représente le nombre de patients appartenant à chacune des modalités à l'intérieur de chaque hôpital.

Certains indices de proximités sont définis pour ce type de tableau de données, citons-en trois:

I. Distance euclidienne

Le premier indice est donné par la distance euclidienne entre deux hôpitaux, chacun étant caractérisé habituellement par un profil de fréquences brutes (Williams [1984]). On définit cette distance comme suit:

$$d_1^2(H_i,H_j) = \sum_{k=1}^{m} (p_{ik}-p_{jk})^2$$

La principale propriété de la distance euclidienne réside dans le fait qu'elle soit géométrique. On n'a qu'à penser au théorème de Pythagore.

II. Distance du khi-carré

Le deuxième indice est donné par la distance du khi-carré entre une paire d'hôpitaux, chacun étant représenté par un profil de fréquences brutes (Benzécri [1973]). Elle s'exprime de la façon suivante:

$$d_2^2(H_i,H_j) = \sum_{k=1}^{m} (p_{..}/p_{.k})[(p_{ik}/p_{i.})-(p_{jk}/p_{j.})]^2$$

où $p_{.k} = \sum_{i=1}^{n} p_{ik}$

$p_{i.} = \sum_{k=1}^{m} p_{ik}$

$p_{..} = \sum_{i=1}^{n} \sum_{k=1}^{m} p_{ik}$

Il est à remarquer que la distance du khi-carré fait intervenir les profils de tous les hôpitaux sous considération. On ne compare donc pas uniquement deux hôpitaux entre eux puisqu'on utilise tous les hôpitaux afin de déterminer $p_{..}$ et $p_{.k}$ qui sont respectivement le nombre total d'hospitalisations dans l'ensemble formé par tous les hôpitaux et le nombre total d'hospitalisations pour chacune des modalités considérées

dans l'ensemble des hôpitaux. Cet indice est à la base de l'analyse des correspondances et il a la propriété d'équivalence distributionnelle i.e. si deux vecteurs-colonne associés à deux modalités différentes sont proportionnels, on peut les sommer en un seul et les distances du khi-carré entre paires d'hôpitaux restent inchangées.

III. Distance informationnelle

Le troisième indice est la distance informationnelle entre deux hôpitaux, chacun étant caractérisé cette fois-ci par un profil de fréquences relatives (Theil [1967]). Elle est définie par:

$$d_3(H_i,H_j) = \sum_{k=1}^{m} p_{ik} \log (p_{ik}/p_{gk}) + \sum_{k=1}^{m} p_{jk} \log (p_{jk}/p_{gk})$$

où $p_{gk} = (p_{ik}+p_{jk})/2$

Si on définit la mesure de la perte d'information sur H_i attribuable à la fusion de H_i dans le groupe formé de H_i et H_j comme étant:

$$PI(H_i,g) = \sum_{k=1}^{m} p_{ik} \log (p_{ik}/p_{gk}).$$

Alors, l'indice de la distance informationnelle peut être interprété de la façon suivante:

Le premier terme du membre de droite de l'équation représente la perte d'information sur H_i attribuable à la fusion de H_i dans le groupe formé de H_i et H_j.

Le deuxième terme représente la perte d'information sur H_j attribuable à la fusion de H_j dans le groupe formé de H_i et H_j.

L'addition de ces deux termes quantifie la perte totale d'information sur les deux hôpitaux du groupe, formé de H_i et H_j.

Dans la prochaine section, nous discuterons du choix de la meilleure mesure de similitude dans l'analyse de regroupement d'hôpitaux.

DISCUSSION

Prenons un exemple simple, soit trois hôpitaux (H_1, H_2, H_3), chacun d'eux ayant un profil de soins caractérisé par deux modalités M_1 et M_2 qu'on peut définir comme étant 2 traitements de nature très différentes i.e. M_1: opérations sur les yeux et M_2: opérations sur la peau.

Tableau 2: Exemple de profils de soins

fréquences brutes

	M_1	M_2
H_1	100	0
H_2	25	25
H_3	0	40

fréquences relatives

	M_1	M_2
H_1	1	0
H_2	0,5	0,5
H_3	0	1

Si on applique les mesures de similitude de la section précédente aux éléments du tableau 2, on obtient le tableau 3 suivant:

Tableau 3: Résultats m-2

	$d_1^2(H_i,H_j)$	$d_2^2(H_i,H_j)$	$d_3(H_i,H_j)$
H_1,H_2	6250	1	0,1874
H_1,H_3	11600	4,44	0,6021
H_2,H_3	850	1	0,1874

Comme on peut le voir, la distance euclidienne mesure une distance plus petite entre H_2 et H_3 qu'entre H_1 et H_2 puisqu'elle tient compte du nombre de patients qui ont subi les interventions dans chacun

des hôpitaux; on rejette donc cette mesure si à cette étape de l'analyse nous ne voulons pas que la taille de l'hôpital influence les résultats.

Si maintenant, on prend toujours trois hôpitaux (H_1, H_2, H_3), mais chacun d'eux ayant un profil de soins caractérisé par trois modalités M_1 ,M_2 et M_3 où M_3 est définie comme étant les opérations sur l'appareil cardiovasculaire. Le tableau 4 ci-dessous illustre une situation hypothétique.

Tableau 4: Exemple de profils de soins

fréquences brutes				fréquences relatives			
	M_1	M_2	M_3		M_1	M_2	M_3
H_1	100	0	0	H_1	1	0	0
H_2	25	25	0	H_2	0,5	0,5	0
H_3	0	0	40	H_3	0	0	1

Pour ce tableau, on ne calcule que les mesures de distance du khi-carré et informationnelle:

Tableau 5: Résultats m=3

	$d_2^2(H_i,H_j)$	$d_3(H_i,H_j)$
H_1,H_2	2,28	0,4316
H_1,H_3	6,27	1,3862
H_2,H_3	7,03	1,3862

Intuitivement, losqu'on regarde le tableau 4 en fréquences relatives, on est porté à croire que l'hôpital H_3 est aussi distant de l'hôpital H_1 qu'il l'est de l'hôpital H_2 mais les calculs faits avec la distance du khi-carré ne réussissent pas totalement à refléter cette perception intuitive; c'est ce qui nous amène à dire que cette distance du khi-carré est inapropriée puisqu'elle tient compte de la taille de chacun des hôpitaux.

Par conséquent, il semble que losqu'on ne veut pas influencer les résultats de l'analyse de regroupement d'hôpitaux par la taille de ceux-ci, il serait préférable d'utiliser la mesure de similitude définie par la distance informationnelle.

Cette dernière mesure fut employée dans les travaux d'un groupe de chercheurs montréalais au début des années 1980 (Lance et Contandriopoulos [1980]) et elle produisit des résultats satisfaisants dans la stratégie de regroupement d'hôpitaux avancée par ceux-ci. Cette discussion portant sur trois mesures de similitudes semble donc justifier, "a postériori", l'utilisation de la distance informationnelle dans le regroupement des hôpitaux québécois en 1980.

REFERENCES

D'AVIGNON, G., DIONNE, A., DUFOUR S., [1989], "Regroupement des centres hospitaliers: une analyse préliminaire.", document de travail no:89-15, LRSA, Faculté des sciences de l'administration, Université Laval.

BENZECRI, J.P., [1973], "Analyse des données, vol.II, l'analyse des correspondances.", Dunod, Paris.

BOURDAGES, J., [1985], "Description du système MED-ECHO, 1982 à 1987.", Ministère de la santé et des services sociaux, Gouvernement du Québec, Québec.

DIONNE, A., [1984], "Notes de cours sur l'analyse multivariée.", Faculté des sciences de l'administration, Université Laval.

DUFOUR, S., [1988], "Analyse des profils de soins des centres hospitaliers du Québec, GRADE RR-88/5, août 88, Faculté des sciences de l'administration, Université Laval.

LANCE, J.M., CONTANDRIOPOULOS, A.P., [1980], " Le regroupement des hôpitaux selon leur production: base d'évaluation de leur performance.", L'Actualité Economique, Vol.56 (2) avril-juin 1980, pp.308-338.

THEIL, H., [1967], "Economics and Information Theory.", Rand-McNally & Co., Chicago.

WILLIAMS, G., [1984], "Linear Algebra with Applications.", Allyn and Bacon, 2-ième édition, p.207 et 253.

EFFET DE L'UTILISATION D'UNE ECHELLE DE REPONSES DICHOTOMIQUE DANS LA SELECTION D'UN VECTEUR DE POIDS

France Lapointe
Université Laval

RESUME

En théorie de la décision, le choix d'un vecteur de poids associés à différents objets selon leur importance relative constitue un problème important. Pour effectuer ce choix, il est commode de comparer les objets deux à deux; c'est ainsi que Saaty [8] suggère d'extraire ce vecteur à partir d'une matrice de comparaisons dont les éléments représentent le degré de préférence exprimée pour un objet par rapport à un autre. Généralement, cette manière de procéder dépend toutefois de l'échelle de réponses employée. Nous examinerons ici le comportement de la méthode de Saaty lorsque l'échelle employée est dichotomique. Quelques exemples illustreront notre propos.

1. INTRODUCTION

1.1 Le problème de la sélection d'un vecteur de poids

On s'intéresse à la comparaison de n objets, $\theta_1, \theta_2, ..., \theta_n$ dans le but de leur accorder des poids $p_1, p_2, ..., p_n \geq 0$ tels que $\sum p_i = 1$ ou simplement de les ranger.

La première situation se rencontre, par exemple, lorsque les n objets considérés constituent différents postes budgétaires à chacun desquels on voudrait associer une part d'un montant global. On parlera alors du problème dit de *l'allocation des ressources*. Dans un contexte statistique, on peut également percevoir p_i comme la probabilité subjective qu'un individu attache à la réalisation future de θ_i. Les différentes méthodes utilisées ici permettront alors de faciliter le choix d'une loi a priori pour $\Theta = \{\theta_1, ..., \theta_n\}$.

On peut toutefois n'être intéressé qu'au rangement des éléments $\theta_1, ..., \theta_n$. C'est le cas, par exemple, lorsque l'on compare les différents candidats à une élection pour un poste de président ou encore lorsque l'objectif premier de la comparaison de certains individus est de déterminer le vainqueur d'un tournoi.

Si la qualité selon laquelle les objets $\theta_1, ..., \theta_n$ sont comparables est mesurable, leur rangement et même la détermination de leurs poids relatifs ne pose aucun problème particulier. Qu'on songe, par exemple, à une situation où on veut ranger un groupe d'individus selon leur taille. Il arrive souvent, par contre, que le critère de comparaison soit difficilement mesurable de façon exacte ou qu'il comporte une certaine part de subjectivité. Le cas des tests de goût constitue un bon exemple d'une telle situation. Il devient alors presqu'impossible à l'être humain de procéder directement à l'attribution des poids p_i à chacun des θ_i ou même au simple rangement des différentes alternatives considérées. De plus, une comparaison simultanée des objets, surtout lorsque n est grand, devient presque impraticable.

Cette dernière considération nous amène à envisager une approche nous permettant d'obtenir malgré tout une comparaison complète des θ_i. On procédera donc à la comparaison des objets deux à deux et ensuite, on colligera cette information dans un tout cohérent. Les comparaisons seront effectuées sur une échelle numérique et on pourra rassembler les informations à l'intérieur d'une matrice dite "de réponses", notée A.

Pour illustrer le procédé, supposons qu'un individu connaisse déjà le vecteur de priorités qu'il désire associer aux éléments de l'ensemble Θ. A la question "Combien de fois préférez-vous θ_i à θ_j?", il serait en mesure de répondre

$$a_{ij} = p_i/p_j \quad \forall\, 1 \le i, j \le n. \tag{1}$$

Ceci permettrait alors de construire une *matrice de réponses* $A = (a_{ij})$ de dimension $n \times n$ à partir de laquelle il serait facile de reconstituer le vecteur de priorités. En particulier, on a

$$p_i \propto a_{ik} \quad \forall\, 1 \le i, k \le n \tag{2}$$

où $(a_{1k}, ..., a_{nk})^t$ est la k-ième colonne de A.

Cependant, remarquons qu'il n'est pas nécessaire de connaître toute la matrice A pour recouvrir le vecteur de priorités. En effet, puisque

$$a_{ji} = 1/a_{ij} \quad \forall\, 1 \le i, j \le n, \tag{3}$$

il suffit de connaître la valeur de a_{ij} pour $j > i$, ce qui réduit (substantiellement) le nombre de questions de n^2 à $n(n-1)/2$. Une matrice qui satisfait la condition (3) est dite réciproque.

Dans le cas où le répondant connaît déjà le vecteur de priorités sous-jacent, on peut encore réduire son fardeau de réponse en remarquant que l'on a

$$a_{ij}\, a_{jk} = a_{ik} \quad \forall\, 1 \le i, j, k \le n. \tag{4}$$

Une matrice qui a la propriété (4) est dite cardinalement cohérente. Si une matrice de réponses possède cette propriété, on peut reconstituer le vecteur de priorités à partir d'un ensemble de $n-1$ éléments judicieusement choisis. Lorsque le répondant connaît les rangs des différents objets, mais pas leurs poids, on devrait alors avoir

$$a_{ij} \underset{(<)}{>} 1,\ a_{jk} \underset{(<)}{>} 1 \ \Rightarrow\ a_{ik} \underset{(<)}{>} 1 \quad \forall\, 1 \le i, j, k \le n. \tag{5}$$

Dans ce cas, la matrice de réponses est dite ordinalement cohérente.

En pratique, toutefois, il arrive fréquemment que la matrice des réponses d'un individu ne soit ni ordinalement, ni cardinalement cohérente. De façon générale, les éléments a_{ij} constituent des estimés des poids relatifs sous-jacents, ce qui s'exprime en posant

$$a_{ij} = \frac{p_i}{p_j}\varepsilon_{ij} \tag{6}$$

où ε_{ij} est l'erreur associée à la comparaison de l'objet i avec l'objet j. Le problème qui se pose est donc le suivant: d'une matrice d'estimations des poids relatifs, on veut reconstituer "au mieux" le vecteur de poids. En d'autres mots, *étant donnée une matrice de réponses A, on cherche une matrice* $\hat{A}$*, positive, réciproque et cohérente qui soit la plus proche possible, en un certain sens, de la matrice de réponses observée et de laquelle on va extraire* $p_1, p_2, \ldots, p_n$ *par la relation (1).*

Ce problème a depuis longtemps suscité l'intérêt des statisticiens. La section suivante nous permettra de tracer un bref historique des travaux effectués sur le sujet.

1.2 Historique

Déjà en 1939, messieurs M.G. Kendall et B. Badington Smith [4] ,s'intéressaient au problème des comparaisons deux à deux dans le contexte du rangement. Plus particulièrement, leurs travaux se sont concentrés sur la définition d'une mesure de l'incohérence ordinale de l'expérimentateur dans ses réponses et sur une façon de colliger les rangements obtenus de "m" répondants à une même expérience.

En 1955, M.G. Kendall [5] aborde encore une fois le problème dans le contexte du rangement et propose une méthode pour extraire, d'une matrice de réponses un peu particulière, le rangement des différents objets considérés. Nous étudierons plus en détail cette proposition dans la section suivante.

Au début des années 1970, le sujet est réexaminé, et ce, de façon indépendante aux travaux de M.G. Kendall et B. Badington Smith. En effet, dans le contexte plus général de l'attribution des poids, T.L. Saaty [8] propose à son tour une méthode pour trouver le vecteur, **p**, de poids.

On retrouve ensuite un certain nombre de publications du même auteur et également de l'un de ses étudiant gradué, L.G. Vargas [10],[11]. En 1984, d'autres chercheurs abordent le problème de l'attribution des poids. Cependant, on remarque qu'alors un effort plus considérable est mis sur la modélisation du problème, celle-ci menant à l'élaboration de méthodes plus à la portée des statisticiens pour extraire le vecteur de poids **p**. Citons ici les travaux de R.E. Jensen [3] et de Piet de Jong [1].

Nous nous attarderons aux méthodes développées par M.G. Kendall [5] et par T.L. Saaty [8]. Nous étudierons plus particulièrement le lien existant entre les solutions qu'ils proposent.

2. UTILISATION D'UNE ECHELLE DE REPONSES DICHOTOMIQUE

Depuis 1977, la méthode proposée par T.L. Saaty pour résoudre le problème de la sélection d'un vecteur de poids a fait l'objet d'un nombre important de publications. Notamment, on discute abondamment de son utilisation dans des domaines aussi variés que le marketing [9], le transport aérien [7]

et le management [2]. Il s'agit en fait de l'une des méthodes les plus à la mode aujourd'hui.

La suggestion de monsieur Saaty prend la forme suivante:

les poids p_i associés aux objets θ_i sont les composantes du vecteur p solution de

$$Ap = \lambda_{max} p \tag{7}$$

tel que $\sum_i p_i = 1$ et où λ_{max} représente la plus grande valeur propre de A.

Selon T.L. Saaty donc, les poids $p_1, p_2, \ldots, p_n$ sont les composantes du vecteur propre normalisé ($\sum p_i = 1$) associé à la valeur propre principale de A. Notons que le théorème de Perron-Frobenius nous assure de l'unicité de cette solution. Mentionnons de plus que puisque Saaty s'est placé dans le contexte de l'attribution des poids, il considère une échelle de réponses qui permet, dans une certaine mesure, l'évaluation de l'écart perçu par le répondant entre deux alternatives. En général, il suggère l'utilisation d'une échelle constituée des entiers de 1 à 9 et de leurs inverses. On retrouve dans [8] une justification de l'emploi d'une telle échelle.

Pour illustrer l'utilisation de cette méthode considérons un exemple:

On désire comparer cinq objets selon leurs poids respectifs en livres. On demande à l'expérimentateur de soupeser une première fois tous les objets pour se faire une idée de leurs poids, ensuite les objets sont comparés deux à deux en les soupesant un après l'autre avec la main droite. Les cinq objets en question sont: une machine à écrire, une radio, une grande valise, un projecteur et une petite valise. L'échelle de réponses employée pour cette expérience est la suivante (voir [8]):

1 les objets i et j sont jugés de poids égaux

2 l'objet i est jugé faiblement plus pesant que l'objet j

⋮

9 l'objet i est jugé beaucoup plus pesant que l'objet j

Les réponses fractionnaires correspondant aux situations où l'objet i est jugé moins pesant que l'objet j.

La matrice de réponses prend ici la forme:

$$A = \begin{bmatrix} 1 & 1/5 & 1/3 & 1/4 & 4 \\ 5 & 1 & 2 & 2 & 8 \\ 3 & 1/2 & 1 & 1/2 & 4 \\ 4 & 1/2 & 2 & 1 & 7 \\ 1/4 & 1/8 & 1/4 & 1/7 & 1 \end{bmatrix}$$

et le vecteur des poids relatif obtenu avec la méthode de Saaty:

$$p = \begin{bmatrix} .09 \\ .40 \\ .18 \\ .29 \\ .04 \end{bmatrix},$$

ce qui est très près de la réalité. En effet, on a pesé ces cinq objets et le vecteur des poids relatifs réels p_r est donné par:

$$p_r = \begin{bmatrix} .10 \\ .39 \\ .20 \\ .27 \\ .04 \end{bmatrix}.$$

M.G. Kendall, pour sa part, s'intéressait plus particulièrement en 1955, au problème du rangement de différentes alternatives. Il suggère dans un premier temps d'obtenir le vecteur p, solution de

$$A^{\bullet}p = \lambda_{max}p \tag{8}$$

tel que $\sum p_i = 1$, et d'ensuite en déduire le rangement des objets considérés. Le rangement est donc déduit de vecteur propre associé à la valeur propre principale de $A^{\bullet}$. Notons que $A^{\bullet}$ n'est pas exactement du même type que les matrices de réponses définies précédemment et cela tient au fait que l'échelle de réponses employée par monsieur Kendall est un peu particulière. Pour bien fixer les idées sur la nature de $A^{\bullet}$, considérons un exemple:

> Dans le cadre d'un tournoi d'échec auquel participaient 6 joueurs, on s'intéresse à déterminer les 3 meilleurs de façon à leur attribuer des prix. Notons que les résultats des différentes rencontres sont codés de la façon suivante:
>
> $$a^{\bullet}_{ij} = \begin{cases} 1 & \textit{si le joueur } i \textit{ a vaincu le joueur } j \\ 0 & \textit{si l'inverse s'est produit} \\ 1/2 & \textit{pour une partie nulle.} \end{cases}$$
>
> Ici,
>
> $$A^{\bullet} = \begin{bmatrix} 1/2 & 1 & 1 & 0 & 1 & 1 \\ 0 & 1/2 & 0 & 1 & 1 & 0 \\ 0 & 1 & 1/2 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1/2 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1/2 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1/2 \end{bmatrix}$$
>
> et le vecteur p solution de (8) est égal à
>
> $$\begin{bmatrix} .28 \\ .12 \\ .23 \\ .13 \\ .12 \\ .12 \end{bmatrix},$$

d'où, le rangement suivant:

$$\begin{bmatrix} 1 \\ 5 \\ 2 \\ 3 \\ 5 \\ 5 \end{bmatrix}.$$

Essentiellement donc, les solutions de M.G. Kendall et de T.L. Saaty sont semblables. Toutefois, les auteurs considèrent le problème sous deux aspects différents: le premier ne s'intéresse qu'au rangement des objets $\theta_1, \theta_2, \ldots, \theta_n$, l'autre, à l'attribution des poids $p_1, p_2, \ldots, p_n$.

Maintenant, que tire-t-on de la méthode de Saaty lorsque l'échelle de réponses employée pour construire A est du type "oui-non" c'est-à-dire du type dichotomique $(a, 1/a)$? On peut concevoir assez facilement que l'emploi d'une telle échelle ne peut vraisemblablement pas nous conduire à l'attribution des poids en tant que tels, les réponses ne nous renseignant pas sur "l'écart" entre deux alternatives. Par contre, le rangement des n objets considérés dans une expérience sera encore possible.

En d'autres mots, il s'agit de savoir comment la méthode de Saaty se comporte lorsqu'elle est utilisée dans le contexte considéré par M.G. Kendall, celui du rangement. De façon plus spécifique, on aimerait savoir:

(1) Si les composantes du vecteur p menant au rangement des θ_i dépendent de la valeur de a.

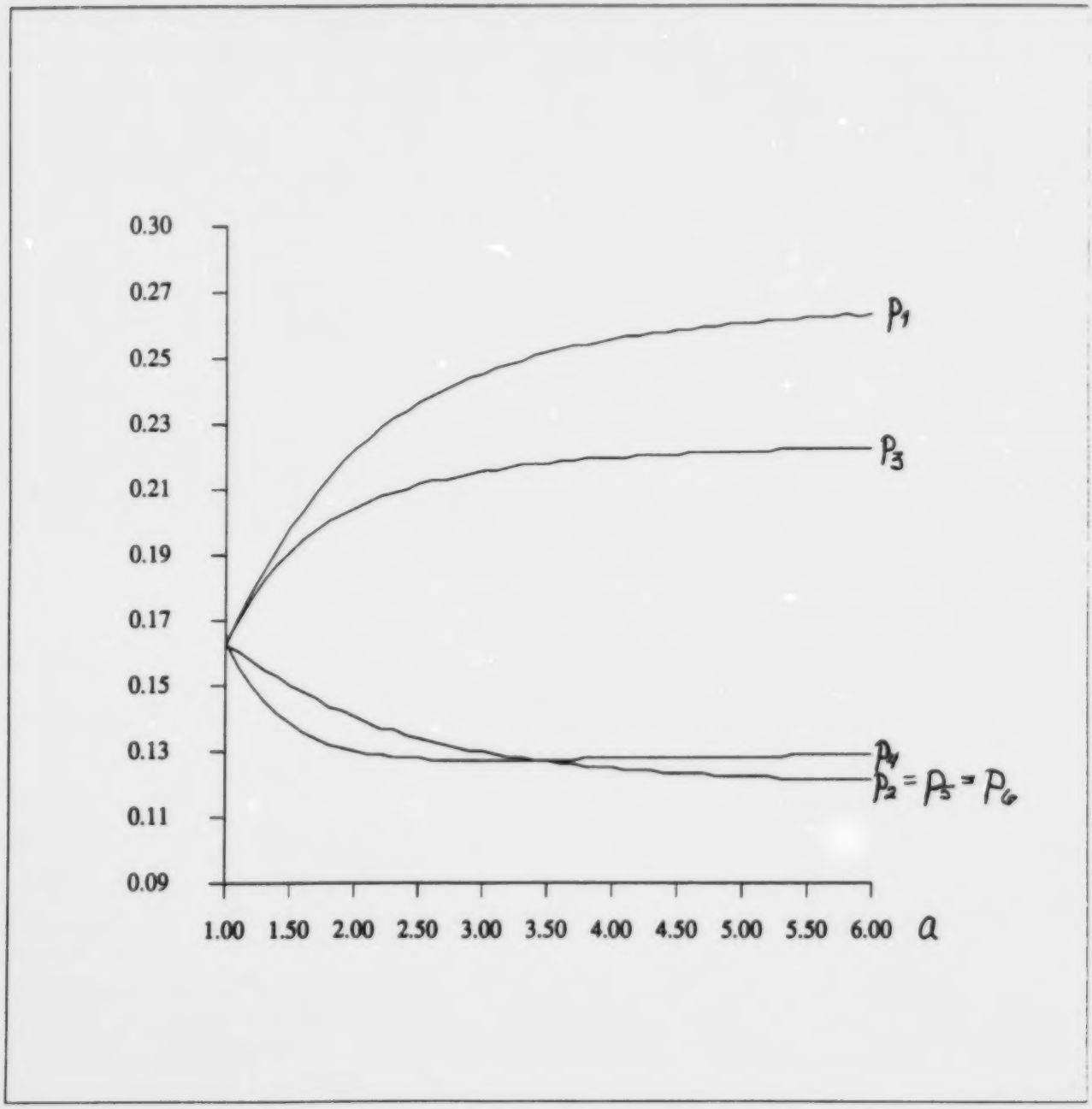
(2) Si la réponse à la question précédente est oui, on voudra savoir si l'ordre induit par ce vecteur de poids est toujours le même quel que soit a.

(3) Finalement, on peut se demander s'il existe, dans le cadre du rangement, un lien plus étroit entre les travaux de M.G. Kendall et ceux de T.L. Saaty.

Considérons à nouveau l'exemple du tournoi d'échec, cette fois-ci toutefois en utilisant une échelle de réponses dichotomique c'est-à-dire qu'on codera les résultats de rencontres de la façon suivante:

$$a_{ij} = \begin{cases} a & \textit{si le joueur } i \textit{ a vaincu le joueur } j \\ 1/a & \textit{si l'inverse s'est produit} \\ 1 & \textit{pour une partie nulle.} \end{cases}$$

Pour différentes valeurs de a allant de 1 à 6, on a obtenu la solution de Saaty. Le graphique qui suit présente les valeurs prises par $p_1, p_2, \ldots, p_6$ en fonction de a.



L'examen de ce graphique nous permet d'illustrer différents résultats. D'abord, il est apparent que les poids, p_i, dépendent de la valeur de a. De plus, le croisement des courbes correspondant à p_2 et p_4 suggère que l'ordre induit n'est pas invariant par rapport à a. Notons de plus que la valeur 1 pour a est à rejeter, celle-ci nous conduisant à l'égalité des poids, et donc des rangs. Finalement, il semble que les p_i tendent vers des valeurs limites lorsque a augmente. Ces valeurs limites sont étroitement liées avec la solution de Kendall. En effet, lorsque a devient suffisamment grand, la solution de Saaty tend vers celle de Kendall. Rappellons que pour cet exemple, la solution de Kendall était la suivante:

de

$$p = \begin{bmatrix} .28 \\ .12 \\ .23 \\ .13 \\ .12 \\ .12 \end{bmatrix},$$

on tirait le rangement r

$$\begin{bmatrix} 1 \\ 5 \\ 2 \\ 3 \\ 5 \\ 5 \end{bmatrix}.$$

Or, lorsque $a = 1000$ par exemple, le vecteur de poids obtenu de la méthode de Saaty est donné par:

$$p = \begin{bmatrix} .28 \\ .12 \\ .23 \\ .13 \\ .12 \\ .12 \end{bmatrix}.$$

On a donc, dans ce cas, concordance des rangs et des poids à deux décimales près.

Cet exemple illustre le comportement général de la méthode de Saaty lorsqu'une échelle de réponses dichotomique est employée. Il existe cependant des situations où ce comportement est un peu différent.

En effet, lorsque la matrice de réponse est de la forme

$$A = \begin{bmatrix} 1 & a & \cdots & a \\ 1/a & 1 & \cdots & a \\ 1/a & 1/a & \cdots & a \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ 1/a & 1/a & \cdots & 1 \end{bmatrix}$$

à un réétiquettage près des objets, on peut montrer que les rangs induits par le vecteur p sont invariants par rapport à a. Cette matrice caractérise une situation où le répondant est parfaitement cohérent au sens ordinal du terme (cf. équation (5)).

3. CONCLUSION

La méthode de Saaty peut nous conduire au rangement de différentes objets $\theta_1, \theta_2, ..., \theta_n$ lorsque l'échelle de réponses employée est dichotomique, c'est-à-dire du type $(a, 1/a)$. De façon générale, par contre, la solution obtenue dépend de la valeur de a. Il semble cependant que le vecteur p duquel le rangement est déduit tende vers un vecter limite lorsque a augmente. Ce dernier correspond au vecteur obtenu de la méthode de Kendall.

Lorsque la matrice de réponses est ordinalement cohérente, on peut montrer que le rangement obtenu de la méthode de Saaty est invariant par rapport à la valeur de a [6].

REFERENCES

[1] de Jong, P. (1984), "A Statistical Approach to Saaty's Scaling Method for Priorites", *Journal of mathematical psychology*, 28, 467-478.

[2] Jensen, R. E. (1982), "Reporting of Management Forecasts - an Eigenvector Model for Elicitation and Review of Forecasts", *Decisions Sciences* , 13, 15-37.

[3] Jensen, R. E. (1984), "An Alternative Scaling Method for Priorities in Hierarchical Structures", *Journal of Mathematical Psychology*, 28 , 317-332.

[4] Kendall, M. G. & Badington Smith, B. (1940), "On the Method of Paired Comparisons", *Biometrika*, 31, 324-345.

[5] Kendall, M. G. (1955), "Further Contributions to the Theory of Paired Comparisons", *Biometrics*, 11, 43-62.

[6] Lapointe, F. (en préparation), "Estimation d'un vecteur de priorités à partir d'une matrice de réponses", *Université Laval.*

[7] NG KYK (1980), "On the Choice of a Military Transport Aircraft Fleet via Worth-Assessement Procedure - A Case-Study", *IEEE Transactions on Systems Man and Cybernetics*, 10 , 894-899.

[8] Saaty, T. L. (1977), "A Scaling Method for Priorities in Hierarchical Structures", *Journal of Mathematical Psychology*, 15, 234-281.

[9] Saaty, T. L. & Wind, Y. (1980), "Marketing Applications of the Analytic Hierarchy Process", *Management Sciences*, 26, 641-658.

[10] Saaty, T. L. & Vargas, L. G. (1984), "Comparison of Eigenvalue, Logarithmic Least Squares and Least Squares Methods in Estimating Ratios", *Mathematical Modelling*, 5, 309-324.

[11] Vargas, L. G. (1983), "Analysis of Sensitivity of Reciprocal Matrices" , *Applied Mathematics and Computation*, 12, 301-320.

La Comparaison Statistique de Classifications Hiérarchiques

François-Joseph Lapointe
Département de Sciences biologiques, Université de Montréal
C. P. 6128, Succursale A, Montréal, Québec H3C 3J7, CANADA

Résumé: Plusieurs mesures de consensus entre classifications ont été développées récemment afin de comparer diverses solutions engendrées par des algorithmes de groupement différents ou basées sur des jeux de données indépendants. Le besoin d'une méthode statistique pour comparer ces consensus n'est plus à démontrer. Certains auteurs (Hubert & Baker 1977 ; Shao & Sokal 1986) ont proposé des tests par permutations ou des simulations afin d'évaluer la signification de certains résultats. Aucune méthode ne peut néanmoins s'appliquer spécifiquement aux phénogrammes utilisés en taxonomie numérique. Nous proposons de tester le consensus entre deux classifications par une double permutation des matrices ultramétriques associées. Chaque matrice produite devra cependant répondre à certains critères définissant une classification aléatoire: topologie, position des feuilles, niveaux de fusion. La permutation des niveaux hiérarchiques des deux classifications suivie d'une seconde permutation des lignes et colonnes de chaque matrice nous permet de générer des classifications aléatoires en accord avec la définition. Le consensus est évalué selon une mesure de distance intermédiaire (Faith & Belbin 1986).

Introduction

La mission originelle du taxonomiste est de tout catégoriser. Pour effectuer ce travail, il développa d'abord de nombreuses méthodes ayant pour fonction de regrouper en classes naturelles tout ce qui s'avérait "classifiable". Le but premier du taxonomiste consistait alors à **classifier**. Cependant, après des années passées à effectuer du groupement, le chercheur a voulu comparer ses résultats à ceux de ses pairs afin de justifier sa recherche. Une nouvelle période s'ammorça: "L'ère de la **comparaison**". Cette nouvelle mission entraîna une révolution méthodologique engendrant une multitude d'indices d'association et de méthodes destinés à comparer des arbres, des hiérarchies, des graphes, des réseaux et des ordinations (Adams 1972; Estabrook, McMorris & Meacham 1985; Margush & McMorris 1981; Neumann 1981; Robinson & Foulds 1981; Rohlf 1982; Stinebrickner 1984; Waterman & Smith 1978; Williams & Clifford 1971). Ces diverses méthodes de comparaison permirent

de répondre à de multiples questions. Deux problèmes majeurs intéressaient cependant les taxonomistes initiés:

1- Est-ce que différentes méthodes de classification produisent des résultats similaires?

2- Est-ce que différentes variables décrivant les mêmes objets taxonomiques entraînent des résultats similaires pour une même méthode?

La théorie des consensus fut développée pour répondre à ces questions. Les indices de consensus, d'une part, ont permis de quantifier la ressemblance entre les classifications alors que les arbres consensus furent développés dans le but de produire des classifications de compromis entre divers arbres non identiques. Néanmoins, malgré les récents développements dans ce domaine, il n'existe encore que très peu de tests permettant de comparer **statistiquement** les classifications. Seules quelques méthodes furent proposées récemment afin de comparer des arbres taxonomiques (Penny & Hendy 1985; Prager & Wilson 1988; Shao & Sokal 1986) ou des cladogrammes biogéographiques (Page 1988; Simberloff, Heck, McCoy & Connor 1981).

Cet article présente une nouvelle méthode de comparaison statistique entre classifications hiérarchiques. Mais, avant d'aller plus loin, il serait utile de définir ce que nous entendons par "classification hiérarchique".

Qu'est-ce qu'une classification hiérarchique?

Le concept de classification, tel qu'accepté en taxonomie est généralement représenté par un arbre. Toutes les propriétés d'une classification taxonomique hiérarchique se retrouvent dans cet arbre mais trois structures retiendront particulièrement notre attention: le tronc, les branches et les feuilles de l'arbre. Chacune de ces composantes biologiques représente un concept mathématique:

-Le tronc de l'arbre, c'est la topologie de la classification, l'architecture du groupement ou encore la structure hiérarchique du dendrogramme.

-Les branches de l'arbre taxonomique, et plus particulièrement la position de celles-ci, décrivent la hauteur des niveaux de fusion de la classification hiérarchique.

-Les feuilles de l'arbre sont les objets que nous classifions.

Un arbre taxonomique au sens strict devra donc se composer d'un tronc, de branches et de feuilles pour mériter le nom de classification hiérarchique ou de dendrogramme. La figure 1 illustre quatre classifications qui diffèrent selon certains aspects de la définition. **Ta** et **Tc** ont des topologies différentes, **Ta** et **Tb** ont des feuilles positionnées différemment, alors que **Tc** et **Td** diffèrent quant à la position de leurs niveaux de fusion. Les dendrogrammes **Tb** et **Td** sont les plus dissemblables car ils diffèrent sur les trois aspects.

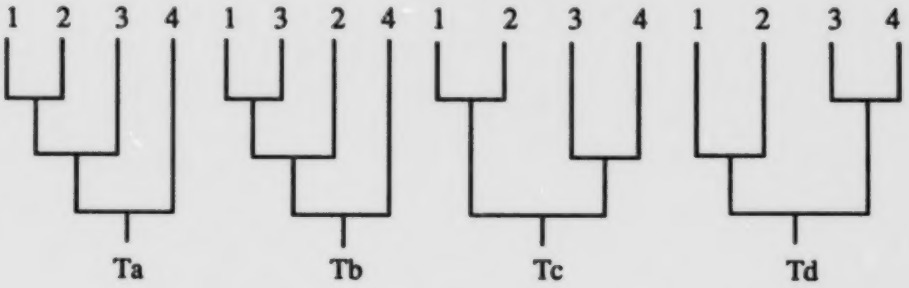


Fig.1. Quatre dendrogrammes qui diffèrent selon divers aspects de la définition de "classification hiérarchique".

Nous verrons que cette définition de classification hiérarchique s'avérera très importante lorsque viendra le temps de comparer divers arbres entre eux.

La formulation des hypothèses

Le développement d'un test statistique se base prioritairement sur la formulation d'une hypothèse que ce test a précisément pour but d'évaluer. Dans notre cas, l'objectif est de tester

la signification de l'association (mesurée par certains indices de consensus) entre deux classifications réelles. Cet objectif consiste opérationnellement à répondre à la question suivante:

Est-ce que les deux arbres taxonomiques que nous comparons se ressemblent plus que n'importe quelle paire d'arbres sélectionnés au hasard dans la forêt ?

Nous pouvons plus précisément reformuler cette question sous la forme d'une hypothèse nulle stipulant ceci: la classification **A** et la classification **B** ne sont pas plus similaires qu'une paire de classifications aléatoires provenant de la même population. Le rejet de cette hypothèse entraînera nécessairement l'acceptation de l'hypothèse alternative affirmant que la classification **A** et la classification **B** sont en effet plus similaires que des classifications sélectionnées au hasard.

Cette approche statistique implique la génération de classifications possédant une structure aléatoire en accord avec la définition de l'arbre taxonomique hiérarchique. Voyons comment ceci s'exprime mathématiquement.

La génération d'arbres aléatoires

Il existe une relation directe et unique entre une classification hiérarchique et la matrice cophénétique contenant les distances ultramétriques entre les objets qui lui sont associés (Sneath & Sokal 1973). Cette bijection mathématique implique une équivalence entre la génération de matrices ultramétriques aléatoires et la simulation d'arbres hiérarchiques (fig. 2). La comparaison de ces matrices représente donc la comparaison des arbres taxonomiques correspondants. Nous désirerons donc, ici, générer des matrices aléatoires acceptables en tant que réalisation de l'hypothèse nulle.

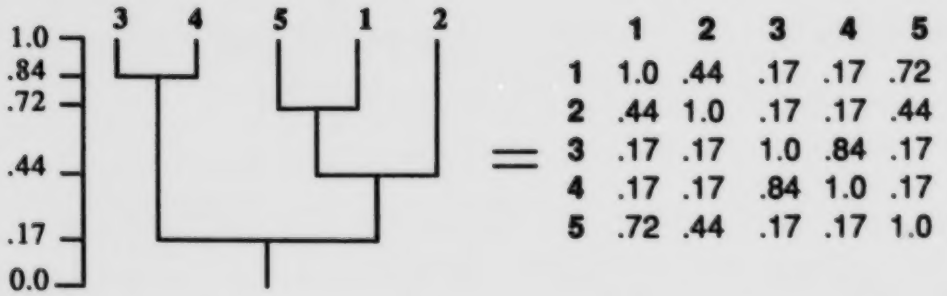


Fig.2. Equivalence entre un dendrogramme et la matrice cophénétique qui lui est associée.

La méthode nous permettant d'effectuer cette opération procède en deux étapes majeures:

1- Génération d'arbres aléatoires non étiquettés

Considérons tout d'abord un vecteur **V** contenant les *(n-1)* niveaux de fusion hiérarchique entre les *n* objets d'un arbre **T**. Il est possible à partir de ce vecteur **V** unique de reconstruire la matrice **A** représentant l'arbre **T** en utilisant la propriété ultramétrique suivante:

$$V_{jk} > \min (V_{ij}; V_{jk}) \qquad \text{où i,j et k } \varepsilon \text{ OTU} \qquad \text{(éq. 1)}$$

et où **OTU** = {1, 2, ..., *(n-1)*, *n* } est l'ensemble des *n* objets de **T**.

Notons que la position des *(n-1)* éléments de ce vecteur **V** dans la matrice **A** décrit la topologie de l'arbre **T** associé à **A**. La permutation de **V** suivi du remplissage de **A** selon la propriété ultramétrique (éq.1) nous permet donc de générer tout **T** possible en fonction de **V**. Cette première étape entraîne la création d'un arbre hiérarchique non étiquetté possédant une topologie aléatoire et des niveaux de fusion placés aléatoirement.

2- Etiquetage aléatoire des topologies générées à l'étape 1

Afin d'obtenir maintenant un arbre étiquetté, il suffit de réadresser aléatoirement les objets de l'ensemble **OTU** aux branches de la topologie de **T**. Cette opération est équivalente

à une simple permutation des lignes et des colonnes de la matrice ultramétrique A tel que décrit par Mantel (1967).

Suite à cette double permutation (du vecteur V premièrement, puis de la matrice A), nous obtenons des arbres aléatoires sous les trois aspects de la définition. Néanmoins, nous n'avons toujours pas de test. Il nous faut maintenant choisir un indice de consensus afin d'évaluer la signification statistique de l'association entre les deux classifications que nous comparons. Lequel choisir parmi la multitude de coefficients possibles?

Le critère de comparaison

Faith (1984) a défini trois types d'indices d'association. Le premier type de coefficients (IC_1) mesure les aspects communs entre les classifications comparées alors que le deuxième groupe d'indices (IC_2) mesure la distance entre deux matrices. Les équations générales représentant ces familles de coefficients sont respectivement:

$$IC_1 = \Sigma\Sigma \min (A_{ij}, B_{ij}) \qquad \text{(éq. 2)}$$

$$IC_2 = \Sigma\Sigma |A_{ij} - B_{ij}| \qquad \text{(éq. 3)}$$

Un troisième patron de consensus décrit par Faith consiste à combiner les deux premier types en un seul coefficient possédant des propriétés uniques. L'Indice de Distance Intermédiaire (IDI) proposé par Faith & Belbin (1986) appartient à cette famille de coefficient d'association. La forme mathématique de cet indice est la suivante:

$$IDI = \Sigma\Sigma [|A_{ij} - B_{ij}| - \min (A_{ij}, B_{ij})] \qquad \text{(éq. 4)}$$

ou bien $$IDI = IC_2 - IC_1$$

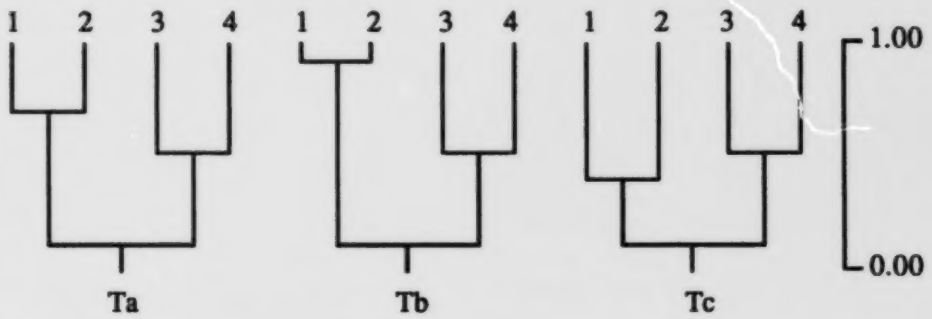
Selon l'indice de consensus que l'on emploiera dans notre test, les résultats pourront varier. Le premier type de coefficients se comporte comme un indice de similarité alors que le

second représente une mesure de distance. L'indice composite de Faith & Belbin est un coefficient-compromis possédant des propriétés intermédiaire tout en étant très informatif. Nous utiliserons donc cet indice comme critère de comparaison pour notre test. Cependant, nous avons préalablement transformé cet Indice de Distance Intermédiaire en un Indice de Similarité Intermédiaire Normalisé afin de standardiser les valeurs du coefficient entre 0 et 1.

$$ISIN = 1 - (\{1 + [(\Sigma\Sigma\ |A_{ij} - B_{ij}| - \min(A_{ij}, B_{ij})) / MAX]\}/2) \qquad (éq.\ 5)$$

où $$MAX = \max[\ \Sigma\Sigma\ |A_{ij} - B_{ij}|,\ \Sigma\Sigma\ \min(A_{ij}, B_{ij})]$$

Selon ce nouvel indice, deux classifications identiques auront toujours un consensus de un. Deux classifications différentes seront néanmoins considérées plus proches si elles possèdent plus de niveaux de fusions élevés en commun (Fig 3).



$ISIN(T_a, T_b) = 0.850.$
$ISIN(T_a, T_c) = 0.786.$

Fig. 3. Comparaison de dendrogrammes selon l'Indice de Similarité Intermédiaire Normalisé

Le test à doubles permutations

Nous avons décrit la procédure nous permettant de générer des arbres hiérarchiques aléatoires puis nous avons choisi le critère de comparaison. Assemblons maintenant toutes les étapes en un test fonctionnel. Cette méthode appartient à la famille des tests par permutations et procède comme suit:

1- On calcule la statistique de référence entre nos deux matrices de départ **A** et **B**.

2- On réalise une double permutation de chaque matrice pour obtenir des classifications **A'** et **B'** aléatoires.

3- On calcule de nouveau le critère statistique entre **A'** et **B'**.

4- On répète les étapes 2 et 3 à plusieurs reprises (le nombre de permutations requises dépend de la taille des matrices comparées).

Cette répétition permet de générer une distribution de référence de la statistique ISIN. La valeur réelle de la statistique mesurée entre les classifications de départ **A** et **B** est testée par rapport à la distribution de référence pour établir la probabilité de cette valeur sous l'hypothèse nulle.

Un programme pour ordinateurs Macintosh écrit en langage Pascal permet de réaliser ce test à doubles-permutations. Ce programme est disponible sur demande auprès de l'auteur.

Conclusion

D'autres tests permettant de comparer statistiquement des classifications furent déjà proposés par divers auteurs. Néanmoins, aucune de ces méthodes ne permet de comparer des arbres en tenant compte simultanément des trois aspects de la définition de classification hiérachique: topologie, position des niveaux de fusion, position des feuilles. La méthode de Shao &Sokal (1986) compare des arbres sans référence aux niveaux de fusion alors que le test par permutations de Hubert & Baker (1977) ne tient pas compte de la topologie des arbres. Le test que nous proposons combine les deux méthodes précédentes. Ce test se veut donc un raffinement des autres méthodes plutôt qu'une procédure statistique visant à remplacer les autres techniques.

Il est à noter cependant que le test que nous venons de décrire devrait être employé uniquement lorsqu'on désire comparer des classification basées sur des données différentes.

La comparaison d'arbres hiérarchiques reconstruits à l'aide de méthode différentes de classification mais se basant sur les mêmes données invalide l'hypothèse nulle de ce test.

Références

Adams, E. N., III. (1972), "Consensus Techniques and the Comparison of Taxonomic Trees," *Systematic Zoology* , 21, 390-397.

Estabrook, G. F., F. R. McMorris & C. Meacham. (1985), "Comparison of Undirected Phylogenetic Trees Based on Subtrees of Four Evolutionary Units," *Systematic Zoology*, 34, 193-200.

Faith, D. P. (1984), "Patterns of Sensitivity of Association Measures in Numerical Taxonomy," *Mathematical Biosciences* , 69, 199-207.

Faith, D. P. & L. Belbin. (1986), "Comparison of Classifications Using Measures Intermediate Between Metric Dissimilarity and Consensus Similarity," *Journal of Classification* , 3, 257-280.

Hubert, L. J. & F. B. Baker. (1977), "The Comparison and Fitting of Given Classification Schemes," *Journal of Mathematical Psychology* , 16, 233-253.

Mantel, N. (1967), "The Detection of Disease Clustering and a Generalized Regression Approach," *Cancer Research* , 27, 209-220.

Margush, T. & F. R. McMorris. (1981), " Consensus *n* -trees," *Bulletin of Mathematical Biology* , 43, 239-244.

Neumann, D. A. (1983), "Faithful Consensus Method for *n* -trees," *Mathematical Biosciences* , 63, 271-287.

Page, R. D. M. (1988), "Quantitative Cladistic Biogeography: Constructing and Comparing Area Cladograms," *Systematic Zoology* , 37, 254-270.

Penny, D. & M. D. Hendy. (1985), "The Use of Tree Comparison Metrics," *Systematic Zoology* , 34, 75-82.

Prager, E. M. & A. C. Wilson. (1988), "Ancient Origin of Lactalbumin from Lysozyme: Analysis of DNA and Amino Acid Sequences," *Journal of Molecular Evolution* , 27, 326-335.

Robinson, D. F. & L. R. Foulds. (1981), "Comparison of Phylogenetic Trees," *Mathematical Biosciences* , 53, 131-147.

Rohlf, F. J. (1982), "Consensus Indices for Comparing Classifications," *Mathematical Biosciences* , 59, 131-144.

Shao, K. & R. R. Sokal. (1986), "Signifiance Tests of Consensus Indices," *Systematic Zoology* , 35, 582-590.

Simberloff, D., K. L. Heck, E. D. McCoy & E. F. Connor. (1981), "There Have Been no Statistical Tests of Cladistics Biogeographical Hypotheses," in *Vicariance Biogeography: A Critique* , eds, G. Nelson & D. E. Rosen, New York: Columbia University Press.

Sneath, P. H. & R. R. Sokal. (1973), " Numerical Taxonomy," W. H. Freeman, San Francisco.

Stinebrickner, R. (1984), "S-consensus Trees and Indices," *Bulletin of Mathematical Biology* , 46, 923-935.

Waterman, M. S. & T. F. Smith. (1978), "On the Similarity of Dendrograms," *Journal of Theoretical Biology* , 73, 789-800.

Williams, W. T. & H. T. Clifford. 1971. On the comparison of two classifications on the same set of elements. *Taxon* , 20:519-522.

Modèles log-linéaires pour des tableaux de contingence complets

Présentation faite au
Colloque sur les applications de la statistique
dans le cadre du
cinquante-septième congrès pour
l'avancement des sciences

le 18 mai 1989
à l'Université du Québec à Montréal

Mylène Lavigne
Statistique Canada

1. INTRODUCTION

Dans le cadre de cette présentation, j'ai donné un bref aperçu de la théorie des modèles log-linéaires pour des tableaux de contingence complets. C'était en quelque sorte un résumé d'une partie de mon mémoire de maitrise que j'ai terminé à l'été 1988. Cet article se veut donc, un compte rendu de cette présentation.

Lorsque l'on veut utiliser la théorie des modèles log-linéaires, c'est qu'on a, en général, recueilli un certain nombre d'observations sur l'état de deux, trois ou plusieurs variables discrètes et on cherche à déterminer s'il existe un lien, une association quelconque entre ces variables discrètes. Les observations sont généralement prises sous trois modèles échantillonnaux: Poisson, Multinômial ou Produit de multinômiales et sont résumées dans un tableau de contingence comme par exemple celui-ci:

Tableau de contingence IxJxK

A	B	C C_1 ... C_K
A_1	B_1 . . . B_J	x_{111} ... x_{11K} . . . x_{1J1} ... x_{1JK}
. . .	. . .	. ...
A_I	B_1 . . . B_J	x_{I11} ... x_{I1K} . . . . . . x_{IJ1} ... x_{IJK}

Ce tableau est en trois dimensions c'est-à-dire, qu'il est composé de trois variables discrètes soient: la variable discrète A avec I niveaux représentés par $A_1,\ldots,A_I$, la variable discrète B avec J niveaux représentés par $B_1,\ldots,B_J$, et la variable discrète C avec K niveaux représentés par $C_1,\ldots,C_K$. Les x_{ijk} (où i=1,...,I, j=1,...,J et k=1,...,K)

représentent les effectifs observés pour chacune des cellules du tableau de contingence.

Pour simplifier la notation, on va tout au long de cet article, considérer qu'on est en présence de seulement trois variables discrètes soient A, B et C avec leurs niveaux respectifs comme dans le tableau précédent. De plus, on va seulement considérer les tableaux complets c'est-à-dire un tableau pour lequel les probabilités théoriques sont toutes positives (en théorie toutes les cellules du tableau sont non vides), par opposition au tableau incomplet, où on sait à priori qu'une ou des cellules du tableau seront vides.

2. LES MODÈLES LOG-LINÉAIRES HIÉRARCHIQUES

Les modèles log-linéaires exposés vont tenter de décrire les associations pouvant exister entre les variables et non pas tenter d'expliquer une variable dépendante à l'aide d'autres variables. Ces modèles log-linéaires sont en quelque sorte analogues au modèle d'analyse de variance mais ils sont appropriés à l'analyse de variables discrètes. Les variables discrètes peuvent être de type dichotomique (par exemple: oui, non), polychotomique quelconque (par exemple: bleu, blanc, rouge), polychotomique ordinale (par exemple: petit, moyen, grand), ou entière (nombre entier naturel).

Donc, par analogie avec les modèles d'analyse de variance, on écrit le modèle log-linéaire le plus général, soit le modèle log-linéaire saturé:

$$\ln p_{ijk} = \theta + \theta_i^A + \theta_j^B + \theta_k^C + \theta_{ij}^{AB} + \theta_{ik}^{AC} + \theta_{jk}^{BC} + \theta_{ijk}^{ABC}$$

pour i=1,...,I, j=1,...,J et k=1,...,K, avec les contraintes suivantes:

$$\sum_{i=1}^{I} \theta_i^A - \sum_{j=1}^{J} \theta_j^B - \sum_{k=1}^{K} \theta_k^C = 0,$$

$$\sum_{i=1}^{I} \theta_{ij}^{AB} - \sum_{j=1}^{J} \theta_{ij}^{AB} - \sum_{i=1}^{I} \theta_{ik}^{AC} - \sum_{k=1}^{K} \theta_{ik}^{AC} - \sum_{j=1}^{J} \theta_{jk}^{BC} - \sum_{k=1}^{K} \theta_{jk}^{BC} = 0$$

et $$\sum_{i=1}^{I} \theta_{ijk}^{ABC} - \sum_{j=1}^{J} \theta_{ijk}^{ABC} - \sum_{k=1}^{K} \theta_{ijk}^{ABC} = 0;$$

Dans ce modèle log-linéaire, la probabilité théorique conjointe que $A=A_i$, $B=B_j$ et $C=C_k$, est représentée par p_{ijk}; les termes d'interaction triple par θ_{ijk}^{ABC}; les termes d'associations partielles entre deux variables étant donné la troisième par θ_{ij}^{AB}, θ_{ik}^{AC}, θ_{jk}^{BC} et l'on a un paramètre associé à chacune des variables.

Ce modèle log-linéaire saturé est construit de façon à ajuster parfaitement les données. Cependant, en pratique on veut trouver un modèle log-linéaire plus simple qui ajustera adéquatement les données et qui nous permettra d'établir les liens qui existent entre les variables. Comme par exemple les modèles suivants:

2.1 Modèle log-linéaire où $\theta_{ijk}^{ABC} = 0$

$$\ln p_{ijk} = \theta + \theta_i^A + \theta_j^B + \theta_k^C + \theta_{ij}^{AB} + \theta_{ik}^{AC} + \theta_{jk}^{BC};$$

2.2 Modèle log-linéaire où $\theta_{ik}^{AC} = \theta_{ijk}^{ABC} = 0$

$$\ln p_{ijk} = \theta + \theta_i^A + \theta_j^B + \theta_k^C + \theta_{ij}^{AB} + \theta_{jk}^{BC};$$

2.3 Modèle log-linéaire où $\theta_{ij}^{AB} = \theta_{ik}^{AC} = \theta_{ijk}^{ABC} = 0$

$$\ln p_{ijk} = \theta + \theta_i^A + \theta_j^B + \theta_k^C + \theta_{jk}^{BC};$$

2.4 Modèle log-linéaire où $\theta^{AB}_{ij} = \theta^{AC}_{ik} = \theta^{BC}_{jk} = \theta^{ABC}_{ijk} = 0$

$$\ln p_{ijk} = \theta + \theta^{A}_{i} + \theta^{B}_{j} + \theta^{C}_{k} .$$

Regardons plus particulièrement le modèle 2.3 où $\theta^{AB}_{ij}=0$, $\theta^{AC}_{ik}=0$ et $\theta^{ABC}_{ijk}=0$. Ce modèle log-linéaire nous indique que les variables A et B sont linéairement indépendantes étant donné C, que les variables A et C sont linéairement indépendantes étant donné B, et que les variables B et C sont linéairement dépendantes étant donné A. On peut faire des remarques analogues pour les autres modèles log-linéaires.

Pour établir un peu de notation, nommons les paramètres associés à une seule variable, les paramètres d'ordre un, les paramètres associés à deux variables, les paramètres d'ordre deux, et ainsi de suite. Donc, par exemple θ^{AB} est un paramètre d'ordre deux. De plus, on nommera le modèle 2.1, le modèle log-linéaire d'ordre deux puisqu'il contient tous les paramètres d'ordre ≤ 2 mais non les paramètres d'ordre supérieur à deux et on nommera le modèle 2.4, le modèle log-linéaire d'ordre un puisqu'il contient tous les paramètres d'ordre un et non les paramètres d'ordre supérieur à un. Donc en général, on nommera un modèle d'ordre Q lorsqu'il contient tous les paramètres d'ordre Q et inférieur à Q, mais non les paramètres d'ordre supérieur à Q.

Il faut aussi dire que tous ces modèles sont des modèles log-linéaires hiérarchiques. Nous donnons le qualicatif hiérarchique à un modèle log-linéaire dans le cas où: si un paramètre associé à un groupe de variables est nul, alors tous les paramètres d'ordre supérieur, associés à ce groupe de variables, sont aussi nuls. Par exemple, un modèle log-linéaire pour un tableau IxJxK avec $\theta^{A}_{i} = 0$ ceci sous-entend que $\theta^{AB}_{ij} = 0$, $\theta^{AC}_{ij} = 0$ et $\theta^{ABC}_{ijk} = 0$.

3. STATISTISQUES D'AJUSTEMENT

On veut obtenir le modèle log-linéaire hiérarchique qui ajuste les données le plus adéquatement possible. Pour vérifier, si le modèle log-linéaire ajuste bien les données, on utilise les deux statistiques d'ajustement suivantes:

$$X^2 = \sum_{ijk} \frac{(x_{ijk} - mp_{ijk})^2}{mp_{ijk}}$$

et

$$G^2 = \sum_{ijk} (x_{ijk}) \ln \frac{x_{ijk}}{mp_{ijk}}$$

où dans les deux cas, la sommation est prise sur toutes les cellules du tableau. La première expression, X^2, est la forme générale pour la statistique du khi-deux de Pearson et la deuxième expression, G^2, est la statistique du rapport de vraisemblance. Si le modèle ajuste bien les données et que la taille de l'échantillon est grande, les deux statistiques sont asymptotiquement distribuées selon une loi khi-deux avec le nombre de degrés de liberté correspondant au modèle log-linéaire (voir Bishop, Fienberg and Holland (1975)). Plusieurs critères ont été donnés pour déterminer si la taille totale de l'échantillon est assez grande afin que la distribution khi-deux donne une bonne approximation de la distribution exacte des statistiques X^2 et G^2 (voir Koehler and Larntz (1980)). Par exemple, un critère acceptable est que l'effectif moyen observé soit d'au moins quatre (voir Larntz (1978)). Ces deux statistiques nous permettent alors de déterminer si le modèle ajuste adéquatement l'ensemble des données.

Cependant, il est intéressant de vérifier si le modèle ajuste adéquatement chacune des observations du tableau. Pour ce faire, on fait une analyse des résidus. Il y a des résidus ajustés et standardisés qui peuvent être calculés et qui nous permettent de faire cette analyse.

4. STRATÉGIES POUR SÉLECTIONNER UN MODÈLE LOG-LINÉAIRE

Avec seulement trois variables, on n'a pas beaucoup de modèles log-linéaires hiérarchiques, on peut donc trouver les statistiques d'ajustement pour chacun des modèles et choisir celui qui ajuste le mieux les données. Mais si on considère seulement une variable de plus, soit quatres variables, on peut calculer 113 modèles log-linéaires hiérarchiques et plus le nombre de variables croit, plus ce nombre de modèles log-linéaires augmente. En pratique, il est impensable de déterminer les statistiques d'ajustement X^2 et G^2 pour tous les modèles log-linéaires hiérarchiques afin de sélectionner le modèle log-linéaire adéquat. Il existe donc des stratégies pour minimiser la procédure de la sélection d'un modèle log-linéaire (voir Benedetti and Brown (1978)).

Le processus de la sélection d'un modèle log-linéaire peut être séparé en deux étapes: la sélection d'un modèle log-linéaire de base (ou de départ) et l'addition ou l'élimination de paramètres θ dans le modèle log-linéaire de base. Les stratégies possibles pour choisir un modèle log-linéaire de base sont des tests simultanés d'ordre Q, des tests d'association de Brown, des tests à l'aide de la statistique Z (Z est le rapport entre l'estimateur d'un paramètre θ et l'estimation de son écart-type) ou encore un modèle log-linéaire de base fixé. Les tests simultanés d'ordre Q nous permettent de vérifier, pour tous les Q possibles, le fait que tous les paramètres d'ordre Q soient simultanément nuls. Le Q maximum, Q^*, pour lequel le test fut significatif, nous indique que nous avons besoin de ces paramètres d'ordre Q^* mais pas nécessairement de tous ces paramètres d'ordre Q^*. Donc un modèle de base peut être le modèle d'ordre Q^* ou Q^*-1. Si le modèle de base est d'ordre Q^*, on tentera d'éliminer des paramètres d'ordre Q^*; si le modèle de base est d'ordre Q^*-1, on ajoutera plutôt des paramètres d'ordre Q^*. Les tests de Brown, nous permettent de déterminer si un paramètre θ à l'intérieur du modèle est significativement différent de zéro. À l'aide des résultats obtenus, on peut trouver un modèle de base. Finalement, on peut utiliser la statistique Z. Lorsqu'on est en présence du modèle log-linéaire saturé et seulement lorsqu'on est

en présence du modèle log-linéaire saturé, cette statistique Z suit asymptotiquement une loi N(0,1) et nous permet alors de déterminer quels sont les paramètres θ significativement différents de zéro dans le modèle log-linéaire saturé. Un modèle de base peut être alors composé de seulement les paramètres θ trouvés significativement différents de zéro.

Lorsqu'on a un modèle log-linéaire de base, on procède à l'élimination ou à l'ajout de paramètres θ. Trois procédures sont souvent utilisées pour ajuster le modèle log-linéaire de base. Ces trois procédures sont: introduction progressive de paramètres θ ("forward"), élimination progressive de paramètres θ ("backward") et pas-à-pas ("stepwise"). La procédure introduction progressive consiste à chaque étape d'ajouter le paramètre θ le plus significatif parmi tous les paramètres θ susceptibles d'être inclus dans le modèle log-linéaire par le principe hiérarchique. La procédure d'élimination progressive consiste à chaque étape d'éliminer le paramètre θ le moins significatif dans le modèle log-linéaire par le principe hiérarchique. La procédure pas à pas suggérée par Goodman (1971) est une combinaison des procédures "forward" et "backward". Au contraire des deux autres procédures, la procédure pas-à-pas permet de réexaminer un paramètre θ que nous avons inclus ou éliminé dans le modèle log-linéaire précédemment. Ces procédures sont les mêmes que celles que nous utilisons pour la régression linéaire multiple.

5. MODÈLES LOG-LINÉAIRES AVEC AU MOINS UNE VARIABLE ORDINALE

Les modèles log-linéaires hiérarchiques ne sont pas nécessairement les plus adéquats lorsque l'une ou plusieurs des variables sont ordinales. Il existe des modèles log-linéaires considérant le fait qu'une ou plusieurs variables soient ordinales. L'objectif de ces modèles log-linéaires est de tenir compte de l'information sur la variable ordinale (voir Agresti (1984)). Voici un exemple de ce genre de modèle log-linéaire:

$$\ln p_{ijk} = \theta + \theta_i^A + \theta_j^B + \theta_k^C + \tau_i^{AB}(v_j - \bar{v}) + \tau_i^{AC}(w_k - \bar{w}) + \beta^{BC}(v_j - \bar{v})(w_k - \bar{w})$$

avec les contraintes:

$$\sum_{i=1}^{I} \theta_i^A = \sum_{j=1}^{J} \theta_j^B = \sum_{k=1}^{K} \theta_k^C = \sum_{i=1}^{I} \tau_i^{AB} = \sum_{i=1}^{I} \tau_i^{AC} = 0.$$

Dans ce modèle log-linéaire, A est une variable discrète quelconque et, B et C sont des variables discrètes ordinales. Les v_j (j=1,...,J) sont les valeurs assignées aux niveaux de la variable B et les w_k (k=1,...,K) sont les valeurs assignées aux niveaux de C. Les moyennes des valeurs assignées aux niveaux des variables B et C sont respectivement v et w.

En général, les valeurs assignées sont à intervalle constant d'un niveau à un autre, sinon il est plus difficile d'interpréter ces modèles log-linéaires. En pratique, on assigne habituellement la valeur 1 au niveau le plus faible, la valeur 2 au second niveau et ainsi de suite. Dans ce cas, il est facile d'interpréter ces modèles log-linéaires. Dans le modèle précédent par exemple, on a β^{BC} qui nous indique l'association uniforme entre B et C et celle-ci est homogène à travers tous les niveaux de A, τ^{AB} qui nous indique l'effet des niveaux de A dans l'association entre A et B et celle-ci est homogène d'un niveau à l'autre de C. Si on avait eu un terme d'interaction triple, les paramètres seraient homogènes à l'intérieur d'un même niveau mais hétérogènes d'un niveau à l'autre.

Ce genre de modèle log-linéaire considérant le fait qu'une ou plusieurs variables soient ordinales est habituellement pris en considération lorsque l'ordre des niveaux d'une variable est pertinent et qu'il est assez simple d'assigner des valeurs au niveau des variables (parce qu'il n'est pas toujours évident d'assigner des valeurs au niveau d'une variable). De plus, les tests associés à ces modèles sont plus puissants que ceux associés aux modèles log-linéaires hiérarchiques.

6. EXEMPLE

Enchainons dès maintenant avec un exemple pour conclure cette discussion. Les données de cet exemple sont résumées dans le tableau de contingence suivant:

Tableau de contingence 4x2x2

M.C.V.*	AGE	FRÉQUENCE DE LA PRATIQUE D'EXERCICE PHYSIQUE			
		NIL	PARFOIS	>1 FOIS/SEM.	CHAQUE JOUR
NON	45-64	88	42	32	74
	≥ 65	45	14	14	48
OUI	45-64	8	6	2	5
	≥ 65	15	4	3	7

*M.C.V.: maladie cardio-vasculaire

Le but de cet exemple est de déterminer s'il existe une association quelconque entre les trois variables discrètes: fréquence de la pratique d'exercice physique, âge et présence d'une maladie cardio-vasculaire. La variable, fréquence de la pratique d'exercice physique, est polychotomique ordinale ayant 4 niveaux: Jamais, Parfois, >1 fois/semaine et tous les jours; la variable, âge, est dichotomique c'est-à-dire qu'elle a deux niveaux: 45-64 ans et ≥ 65 ans; et la troisième variable, présence d'une maladie cardio-vasculaire, a aussi deux niveaux: non et oui.

Pour établir s'il existe une association quelconque entre les trois variables considérons d'abord le modèle log-linéaire saturé (le modèle log-linéaire le plus général) qui ajuste parfaitement les données. Puisqu'on est en présence du modèle log-linéaire saturé, calculons la

statistique Z pour chacun des paramètres θ. Les résultats sont les suivants:

$$Z_1^F = 5,096; \quad Z_2^F = -1,801; \quad Z_3^F = -3,432; \quad Z_4^F = 2,028;$$

$$Z_1^A = 1,455; \quad Z_2^A = -1,455; \quad Z_1^M = 10,885; \quad Z_2^M = -10,885;$$

$$Z_{11}^{FA} = -1,022; \quad Z_{21}^{FA} = 1,586; \quad Z_{31}^{FA} = -0,129; \quad Z_{41}^{FA} = -0,786;$$

$$Z_{12}^{FA} = 1,022; \quad Z_{22}^{FA} = -1,586; \quad Z_{32}^{FA} = 0,129; \quad Z_{42}^{FA} = 0,678;$$

$$Z_{11}^{FM} = -0,850; \quad Z_{21}^{FM} = -1,116; \quad Z_{31}^{FM} = 0,347; \quad Z_{41}^{FM} = 1,219;$$

$$Z_{12}^{FM} = 0,850; \quad Z_{22}^{FM} = 1,116; \quad Z_{32}^{FM} = -0,347; \quad Z_{42}^{FM} = -1,219;$$

$$Z_{11}^{AM} = 2,802; \quad Z_{12}^{AM} = -2,802 \quad Z_{21}^{AM} = -2,802; \quad Z_{22}^{AM} = 2,802;$$

$$Z_{111}^{FAM} = 0,556; \quad Z_{211}^{FAM} = -0,501; \quad Z_{311}^{FAM} = 0,286; \quad Z_{411}^{FAM} = -0,340;$$

$$Z_{121}^{FAM} = -0,556; \quad Z_{221}^{FAM} = 0,501; \quad Z_{321}^{FAM} = -0,286 \quad Z_{421}^{FAM} = 0,340;$$

$$Z_{112}^{FAM} = -0,556; \quad Z_{212}^{FAM} = 0,501; \quad Z_{312}^{FAM} = -0,286; \quad Z_{412}^{FAM} = 0,340;$$

$$Z_{122}^{FAM} = 0,556; \quad Z_{222}^{FAM} = -0,501; \quad Z_{322}^{FAM} = 0,286; \quad Z_{422}^{FAM} = -0,340;$$

où Z est le rapport entre l'estimation du paramètre θ et l'estimation de son écart-type. Cette statistique Z suit asymptotiquement une loi N(0,1) (puisqu'on est en présence du modèle log-linéaire saturé) et on peut donc comparer Z à la valeur 1,96 pour un test à un niveau de 5%. On se rend compte alors que seulement les paramètres associés à la variable F: fréquence de la pratique d'exercice physique, les paramètres associés à la variable M: Présence de maladie cardio-vasculaire et les paramètres associés aux variables AM: Age et présence de maladie cardio-vasculaire sont significativement différents de zéro.

Donc, un modèle de base approprié est le modèle contenant tous les paramètres d'ordre un et le paramètre d'ordre deux associé aux variables Age et présence de maladie cardio-vasculaire, c'est-à-dire le modèle log-linéaire suivant:

$$\ln p_{ijk} = \theta + \theta_i^F + \theta_j^A + \theta_k^M + \theta_{jk}^{AM};$$

Cependant, on peut aussi sélectionner un modèle de base par les tests simultanés d'ordre Q. On obtient les résultats suivant pour Q égale un, deux et trois:

Q	D.L.	G^2	P
3	3	0,67	0,88140
2	7	18,30	0,01069*
1	5	361,95	0,00000*

Pour Q=3, on teste H_o: tous les paramètres d'ordre trois sont simultanément nuls; pour Q=2, on teste H_o: tous les paramètres d'ordre deux sont simultanément nuls; et pour Q=1, on teste H_o: tous les paramètres d'ordre un sont simultanément nuls. On constate (toujours au niveau 5%) que le Q maximum pour lequel le test est significatif est: Q=2. Ce qui signifie qu'on a besoin de paramètres d'ordre deux mais pas nécessairement tous les paramètres d'ordre deux. Donc, le modèle de base peut être le modèle d'ordre deux ou le modèle d'ordre un. Si on considère le modèle d'ordre deux, on essayera d'éliminer des paramètres d'ordre deux et si on considère le modèle d'ordre un, on tentera d'ajouter des paramètres d'ordre deux. On aurait pu aussi effectuer les tests de Brown, dans ce cas, on trouve le même modèle de base que l'on a trouvé avec la statistique Z.

A partir des modèles de base, on a utilisé la méthode "pas à pas" pour trouver le modèle qui ajustait le plus adéquatement les données et on a trouvé le modèle suivant:

$$\ln p_{ijk} = \theta + \theta_i^{F} + \theta_j^{A} + \theta_k^{H} + \theta_{jk}^{AH} ;$$

A l'aide des deux statistiques X^2 et G^2 (X^2=8,52 avec P=0,482 et G^2=8,27 avec P=0,507), on constate que ce modèle ajuste adéquatement les données au niveau 5%. Et si on fait une analyse des résidus, on trouve aussi que le modèle est adéquat pour chacune des observations.

Cependant, on aurait pu considérer les modèles log-linéaires considérant le fait que la variable, Fréquence de la pratique de l'exercice, est ordinale. Considérons par exemple, le modèle suivant:

$$\ln p_{ijk} = \theta + \theta_i^F + \theta_j^A + \theta_k^M + \tau_j^{FA}(u_i - \bar{u}) + \tau_k^{FM}(u_i - \bar{u}) + \theta_{jk}^{AM} + \tau_{jk}^{FAM}(u_i - \bar{u}).$$

Si on teste, H_o : ce modèle est adéquat (P=0,564), on ne rejette pas cette hypothèse et si on teste, H_o : $\tau^{FAM} = 0$ ($P > 0,3$), on ne rejette pas non plus cette hypothèse. On peut donc considérer maintenant ce modèle log-linéaire, en excluant le paramètre τ^{FAM}, c'est-à-dire le modèle suivant:

$$\ln p_{ijk} = \theta + \theta_i^F + \theta_j^A + \theta_k^M + \tau_j^{FA}(u_i - \bar{u}) + \tau_k^{FM}(u_i - \bar{u}) + \theta_{jk}^{AM}.$$

Mais, si on teste les hypothèses H_o: $\tau^{FA} = 0$ (P=0,500), H_o: $\tau^{FM} = 0$ (P=0,098) et H_o: $\theta^{AM} = 0$ (P=0,001) à partir de ce modèle, on constate que le P, le plus élevé, pour lequel le test n'est pas significatif est celui associé à H_o: $\tau^{FA} = 0$. On l'élimine donc du modèle et on se retrouve avec le modèle:

$$\ln p_{ijk} = \theta + \theta_i^F + \theta_j^A + \theta_k^M + \tau_j^{FA}(u_i - \bar{u}) + \tau_k^{FM}(u_i - \bar{u}) + \theta_{jk}^{AM}.$$

On teste à nouveau les hypothèses H_o: $\tau^{FM} = 0$ (P=0,100) et H_o: $\theta^{AM} = 0$ (P=0,001). Puisqu'on trouve que τ^{FM} n'est pas significativement différent de zéro, on peut aussi l'éliminer du modèle. On se retrouve donc avec le modèle log-linéaire hiérarchique qu'on a trouvé avant de considérer que la variable, Fréquence de la pratique d'exercice physique, était ordinale. Donc, le fait que la variable, fréquence de la pratique d'exercice physique, soit ordinale n'apporte pas de différence quant au choix du modèle log-linéaire.

On conclut donc que la variable, Fréquence de la pratique de l'exercice physique est linéairement indépendante des deux autres variables et que les variables, âge et la présence d'une maladie cardio-vasculaire, sont linéairement dépendantes.

BIBLIOGRPHIE

AGRESTI, A. (1984). Analysis of ordinal categorical data. Wiley, New-york.

AITKIN, M. (1983). Linear statistical analysis of discrete data. Wiley, New-york.

BENEDETTI, J.K., BROWN, M.B. (1978). Strategies for the selection of log-linear models. Biometrics, 34, 680-686.

BROWN, M.B. (1976). Screening in multidimensional contingency tables. Applied Statistics, 25, 37-46.

BISHOP, Y.M.M., FIENBERG, S.E. AND HOLLAND. P.W. (1975). Discrete Multivariate Analysis. MIT Press, Cambridge, Mass.

FIENBERG, S.E. (1980). The analysis of cross-classified categorical data. The MIT Press, Cambridge, Mass.

GOODMAN, L.A. (1971). The analysis of multidimensional contingency tables: stepwise procedures and direct estimation methods for building models for multiple classifications. Technometrics, 13, 33-61.

GOODMAN, L.A. (1984). The analysis of cross-classified data having ordered categoricals. Havard University Press, Cambridge.

KOEHLER, K., and LARNTZ, K. (1980). An emperical investigation of goodness-of-fit statistics for sparse multinomials. Journal of American Statistical Assoc. 75, 336-344.

LARNTZ, K. (1978). Small-Sample comparaison of exact levels for chi-squared goodness-of-fit statistics. Journal of American Statistical Assoc., 73, 253-263.

Estimation de la Taille d'une Population de Saumonneaux à l'Aide d'un Modèle de Capture-Recapture Avec Stratification

Nathalie Plante et Louis-Paul Rivest
Université Laval

RESUME

Dans le cadre d'un modèle de capture-recapture, l'estimateur de Petersen est couramment utilisé pour évaluer la taille d'une population animale. Il peut cependant être un mauvais estimateur lorsque les hypothèses sous-jacentes au modèle sont violées. En particulier, les probabilités de capture varient parfois dans le temps ou l'espace. On a alors recours à une stratification du temps ou de l'espace menant à un estimateur plus adéquat. Ces résultats sont illustrés à l'aide de données de la dévalaison de la rivière Conne (Terre-Neuve) par une population de saumonneaux.

1. INTRODUCTION

Un des premiers pas dans l'étude d'une population animale sauvage est la détermination de sa taille. Celle-ci peut être estimée à l'aide de méthodes de capture-recapture, ce qui permet d'éviter un dénombrement complet de la population.

Le Ministère du Loisir, de la Chasse et de la Pêche effectue des expériences de capture-recapture depuis 1983 sur la rivière de la Trinité (Baie-Trinité) et la rivière St-Jean (Gaspé). L'objectif poursuivi est d'estimer la taille des populations de saumonneaux qui dévalent ces rivières vers la mer, au printemps. Comme on peut dénombrer les saumons qui reviennent frayer à leur rivière natale, ces estimations permettent de développer un système de prédiction des retours de saumons adultes en eau douce. On peut alors déterminer des niveaux de récoltes favorisant la survie du saumon atlantique (Caron et Mercier, 1987).

Des travaux semblables ont également été effectués en 1987 par Pêches et Océans Canada sur la rivière Conne, à Terre-Neuve (Dempson et Stansbury, à paraître).

2. EXPERIENCE DE CAPTURE-RECAPTURE

Réduite à sa plus simple expression, l'expérience consiste en deux échantillonnages successifs. Les animaux capturés dans le premier échantillon sont d'abord marqués, puis rendus à la population. Après un certain temps durant lequel les animaux marqués et non marqués se mélangent, on tire un second échantillon comprenant un certain nombre d'animaux marqués et non marqués (figure 1).

Figure 1. Expérience de capture-recapture.

Ce sont des expériences de ce genre qui ont été effectuées par le Ministère du Loisir, de la Chasse et de la Pêche et par Pêches et Océans Canada. Pour ce faire, des trappes ont été installées en deux zones distinctes de la rivière à l'étude: une zone de capture et une zone de recapture en aval de la précédente. Chaque jour, les saumonneaux capturés dans la première zone ont été marqués, puis remis à l'eau au même endroit. L'espacement entre les deux zones a ensuite permis le mélange des poissons marqués et non marqués, dont un certain nombre ont été capturés dans la zone de recapture.

3. ESTIMATEUR DE PETERSEN

Les données recueillies à partir d'une expérience de capture-recapture se résument aisément à l'aide d'un tableau de fréquences 2×2 incomplet:

Tableau 1. Fréquences observées

1er échant.	2e échant. présence	absence	
présence	n_{11}	n_{12}	$n_{1.}$
absence	n_{21}		
	$n_{.1}$		

$n_{1.}$ est le nombre d'animaux dans le premier échantillon, $n_{.1}$ le nombre dans le second et n_{11} le nombre d'animaux présents dans les deux échantillons. Le nombre d'animaux observés dans le premier échantillon seulement est $n_{12} = n_{1.} - n_{11}$, alors qu'il y en a $n_{21} = n_{.1} - n_{11}$ uniquement dans le second. Quant à la cellule vide, elle représente les animaux n'ayant jamais été capturés.

Si l'on suppose que la proportion d'animaux marqués dans le second échantillon est un estimateur raisonnable de la proportion inconnue d'animaux marqués parmi la population, l'estimateur $\hat{N}_p$ de N, la taille de la population, est tel que

$$\frac{n_{11}}{n_{.1}} = \frac{n_{1.}}{\hat{N}_p}, \quad ou \quad \hat{N}_p = \frac{n_{1.}n_{.1}}{n_{11}} .$$

Il s'agit de l'estimateur de Petersen couramment utilisé dans l'étude des populations animales. Or cet estimateur n'est adéquat que s'il n'a tendance ni à sous-estimer, ni à surestimer la taille de la population, c'est-à-dire s'il est convergent. Il faut pour cela que les probabilités de capture (1er échantillon) et les probabilités de recapture (2ème échantillon) soient non corrélées. En particulier, il suffit que tous les animaux aient la même probabilité de faire partie du premier échantillon (égalité des probabilités de capture) ou du second échantillon (égalité des probabilités de recapture) pour que l'estimateur de Petersen soit adéquat (Seber, 1982).

Pour ce qui est des saumonneaux, puisque la dévalaison se poursuit sur une période de temps relativement longue (environ 1 mois), il serait hasardeux de considérer que les probabilités de capture ou recapture sont constantes. En effet, divers facteurs tels que les déplacements ou ajouts de trappes, variations du niveau de l'eau ou des conditions climatiques, peuvent modifier ces probabilités dans le temps. D'autre part, il est possible qu'un saumonneau soit plus facile à capturer s'il passe d'un certain côté de la rivière, par exemple. Il peut donc y avoir aussi hétérogénéité des probabilités de capture dans l'espace.

L'exemple qui suit présente une situation où l'estimateur de Petersen a tendance à sous-estimer la taille réelle de la population, dû à une corrélation positive (égale à 1) entre les probabilités de capture et de recapture. Supposons qu'une population de saumonneaux puisse être divisée en trois groupes selon leur moment de passage aux zones de capture et recapture, les probabilités de capture et recapture étant homogènes dans chacun des groupes.

Tableau 2. Exemple d'estimation biaisée

	PERIODES DE TEMPS			
	1	2	3	total
nombre saum.	2500	5000	2500	10000 = N
E(saum. marqués)	250	1500	500	2250 = E($n_{1.}$)
$p_{capture}$	.10	.30	.20	
E(saum. ds 2e éch.)	250	3500	1000	4750 = E($n_{.1}$)
$p_{recapture}$	.10	.70	.40	
E(saum. recapturés)	25	1050	200	1275 = E(n_{11})

Aux 2500 premiers saumonneaux dévalant la rivière sont associées des probabilités de capture et recapture de .10 et .10. Selon toute vraisemblance, on devrait en capturer environ 250 (2500×.10) dans le premier échantillon (que l'on marque), et autant dans le second. Puisque 10% des saumonneaux ont été marqués, le second échantillon devrait contenir 25 (250×.10) saumonneaux marqués, c'est-à-dire ayant déjà été capturés. Aux 5000 saumonneaux suivants sont associées d'autres probabilités de capture et recapture, et ainsi de suite. On devrait donc avoir un total d'environ 2250 saumonneaux à la capture et

4750 à la recapture, dont 1275 ayant déjà été capturés. On a donc $E(n_{1.})E(n_{.1})/E(n_{11}) = 8382 < 10000 = N$, c'est-à-dire que l'estimateur de Petersen présente un biais négatif pour la taille de la population.

Pour remédier au problème d'hétérogénéité des probabilités de capture et de recapture, on peut envisager un modèle plus général avec stratification.

4. MODELE AVEC STRATIFICATION

Plaçons-nous dans le cadre d'une expérience de capture-recapture où il n'y a ni naissance, ni immigration et supposons que les "mouvements" des animaux sont indépendants et homogènes pour les marqués et non-marqués (en termes de déplacements, vulnérabilité à la capture, mortalité ou émigration). Nous supposerons aussi dorénavant que les probabilités de capture varient dans le temps. L'expérience de capture-recapture est schématisée à la figure 2.

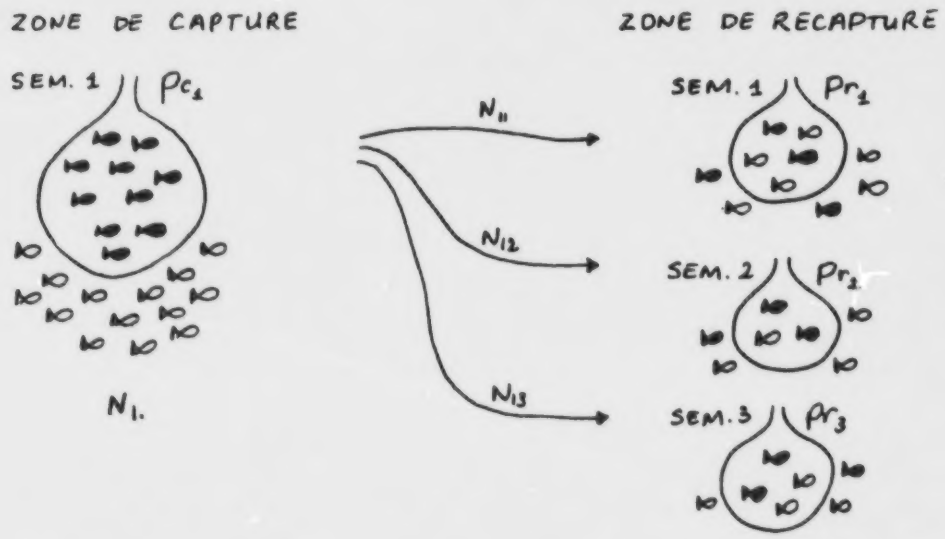


Figure 2. Expérience de capture-recapture pour le modèle avec stratification.

$N_{1.}$ est le nombre d'animaux étant passés à la zone de capture pendant la semaine 1. De ce nombre, N_{11} sont passés à la zone de recapture à la semaine 1, N_{12} à la semaine 2 et N_{13} à la semaine 3. Comme l'indique la figure 2, à chaque semaine sont associées une probabilité de capture (p_c) et une probabilité de recapture (p_r). De plus, les animaux de strates de capture différentes sont

marqués de façons distinctes.

Chapman et Junge (1956), puis Darroch (1961), ont développé un modèle pour les cas où il y a un nombre égal de strates de capture et de recapture. Nous supposerons ici qu'il y a s strates de capture ($1 \leq i \leq s$) et t strates de recapture ($1 \leq j \leq t$). Les observations peuvent alors être résumées à l'aide du tableau suivant:

Tableau 3. Fréquences observées

ZONE DE CAPTURE		ZONE DE RECAPTURE présence 1	2	...	t	absence	total
	1	m_{11}	m_{12}	...	m_{1t}	a_1	c_1
	2		m_{22}	...	m_{2t}	a_2	c_2
présence	...				...	...	...
			0				
	s				m_{st}	a_s	c_s
absence		b_1	b_2	...	b_t		
total		r_1	r_2	...	r_t		

Notons que l'on est en présence d'un tableau (s+1)×(t+1) dont une cellule, celle correspondant aux animaux qui ne sont jamais capturés, est vide. On doit donc estimer N, la taille de la population, à l'aide des seules fréquences observées. D'autre part, de manière à obtenir une estimation qui soit la plus précise possible, il serait utile de regrouper les strates dont les probabilités de capture ou recapture sont homogènes.

Dans ce but, exprimons les valeurs prédites des variables aléatoires du tableau précédent en fonction des paramètres suivants:

N_{ij} = nombre d'animaux dans la strate i au moment de la capture et dans la strate j au moment de la recapture ($N = \sum_{i=1}^{s}\sum_{j=1}^{t} N_{ij}$),

p_{c_i} = probabilité de capture pour un animal dans la strate i,

p_{r_j} = probabilité de recapture pour un animal dans la strate j.

On obtient alors le tableau suivant pour les valeurs prédites:

Tableau 4. Valeurs prédites

		RECAPTURE			
		présence			absence
CAPTURE		1	...	t	
	1	$p_{c_1} N_{11} p_{r_1}$	...	$p_{c_1} N_{1t} p_{r_t}$	$p_{c_1} \sum_j N_{1j} (1-p_{r_j})$
présence	...		...	...	...
	s	0		$p_{c_s} N_{st} p_{r_t}$	$p_{c_s} \sum_j N_{sj} (1-p_{r_j})$
absence		$\sum_i (1-p_{c_i}) N_{i1} p_{r_1}$	...	$\sum_i (1-p_{c_i}) N_{it} p_{r_t}$	

Une reparamétrisation permet de simplifier l'écriture des valeurs prédites:

$$\theta_{ij} = p_{c_i} N_{ij} p_{r_j} \quad , \quad \textit{pour tout } (i,j) \, ,$$

$$\beta_i = \frac{1-p_{c_i}}{p_{c_i}} \quad , \quad \textit{pour tout } i \, ,$$

$$\alpha_j = \frac{1-p_{r_j}}{p_{r_j}} \quad , \quad \textit{pour tout } j \, .$$

Le tableau des valeurs prédites devient:

Tableau 5. Valeurs prédites après reparamétrisation

		RECAPTURE			
		présence			absence
CAPTURE		1	...	t	
	1	θ_{11}	...	θ_{1t}	$\sum_j \alpha_j \theta_{1j}$
présence	...		...	...	...
	s	0		θ_{st}	$\sum_j \alpha_j \theta_{sj}$
absence		$\sum_i \beta_i \theta_{i1}$	...	$\sum_i \beta_i \theta_{it}$	

Posons maintenant

$$\gamma_i = \sum_j \alpha_j \theta_{ij}, \quad \textit{pour tout } i \,,$$

$$\psi_j = \sum_i \beta_i \theta_{ij}, \quad \textit{pour tout } j \,.$$

Le paramètre à estimer, $N = \sum_i \sum_j N_{ij}$, peut alors s'écrire en fonction de la nouvelle paramétrisation:

$$N = \sum_i \sum_j (\beta_i+1)\, \theta_{ij}\, (\alpha_j+1)$$

ou

$$N = \sum \textit{cellules non vides du tableau } 5 + \begin{cases} \sum_i \beta_i \gamma_i \\ \sum_j \alpha_j \psi_j \end{cases}.$$

On remarque que les quantités $\sum \beta_i \gamma_i$ et $\sum \alpha_j \psi_j$ sont en fait égales à la valeur prédite pour la cellule vide du tableau précédent. Il apparaît ainsi que pour estimer N, on doit calculer soit une estimation des β_i, soit une estimation des α_j. Pour ce faire, on peut maximiser la fonction de vraisemblance multinomiale simple du tableau (s+1)×(t+1). Or la maximisation simultanée par rapport à N et aux paramètres du tableau 5 est complexe. On peut simplifier la procédure à l'aide d'une factorisation de la fonction de vraisemblance. Cela permet d'estimer dans un premier temps les paramètres du tableau 5 de façon conditionnelle, c'est-à-dire en maximisant la fonction de vraisemblance multinomiale du tableau formé des (s+1)×(t+1)-1 cellules non vides (la somme des fréquences observées étant fixée). N est ensuite estimée par $\hat{N}_c$ qui maximise la vraisemblance de la cellule vide et qui est fonction des estimations précédentes (Bishop, Fienberg et Holland, 1975).

Lorsqu'il y a autant de strates de capture que de recapture (s=t), l'estimateur $\hat{N}_c$ et sa variance asymptotique sont équivalents à ceux obtenus par Darroch(1961). Dans le cas contraire, le modèle proposé ici permet l'estimation des paramètres à l'aide d'une procédure itérative. On obtient également une forme générale pour la variance asymptotique de $\hat{N}_c$ (Plante, en préparation).

5. EXEMPLE D'APPLICATION

Les données analysées proviennent de la rivière Conne (Terre-Neuve) et ont été recueillies par Pêches et Océans (Dempson et Stansbury, à paraître). Chaque strate correspond à une période de sept jours.

Tableau 6. Fréquences observées

		RECAPTURE							
		présence						absence	total
CAPTURE		1	2	3	4	5	6		
	1	149	126					1561	1836
	2		308	65				1235	1608
	3			161	77			884	1122
présence	4				67	7		215	289
	5					17	3	71	91
	6						13	16	29
absence		1895	8503	2184	525	155	118		
total		2044	8937	2410	669	179	134		

D'abord, des tests permettent de rejeter les hypothèses d'égalité des probabilités de capture et des probabilités de recapture. Ceci indique que l'estimateur de Petersen pour la taille de la population risque d'être inadéquat. On doit donc recourir à une estimation dérivée du modèle avec stratification. L'ajustement de ce modèle au tableau complet fournit les estimations suivantes:

Tableau 7. Estimations des paramètres pour le tableau complet

semaines	$\hat{p}_{r_j}$	$\hat{\alpha}_j$
1	.11	7.8
2	.24	3.1
3	.19	4.2
4	.26	2.8
5	.20	4.0
6	.45	1.2

$$\hat{N}_c = 71\ 127, \qquad \hat{V}(\hat{N}_c) = 5\ 043\ 193.$$

Des tests de seuil 5% sur les paramètres α_j pris deux à deux ne permettent pas de détecter de différence significative entre les probabilités de recapture des strates 2, 3, 4 et 5. Après regroupement des 4 colonnes correspondantes, on obtient les estimations suivantes:

Tableau 8. Estimations des paramètres pour le tableau regroupé

semaines	$\hat{p}_{r_j}$	$\hat{\alpha}_j$
1	.12	7.6
2 à 5	.23	3.4
6	.44	1.3

$$\hat{N}_c = 71\ 752, \qquad \hat{V}(\hat{N}_c) = 4\ 228\ 977.$$

Les tests sur les paramètres α_j ne permettent plus de faire d'autres regroupements de strates de recapture. On obtient finalement un intervalle de confiance conditionnel pour N (à 95%):

[67 722, 75 783].

De manière à valider la procédure proposée, on a pu calculer une estimation inconditionnelle pour N et l'intervalle de confiance correspondant (à 95%):

$$\hat{N} = 71\ 742,$$

[67 897, 75 977].

On remarque que cet intervalle est très semblable au précédent.

Finalement, l'estimation de N selon Petersen fournit les résultats suivants:

$$\hat{N}_p = 71\ 956, \qquad \hat{V}(\hat{N}_p) = 3\ 876\ 319,$$

[68 097, 75 815].

La faible différence entre $\hat{N}_p$ et $\hat{N}$ ou $\hat{N}_c$ peut être attribuable au fait que la procédure nous ait permis de regrouper les strates de recapture 2 à 5 qui représentent à la majeure partie des observations. D'autre part, il peut arriver que la corrélation entre les probabilités de capture et de recapture soit négligeable. Dans ce cas, l'estimateur de Petersen est adéquat, bien que ni les probabilités de capture, ni celles de recapture ne soient constantes.

REFERENCES

Bishop, Y. M. M., Fienberg, S. E. & Holland, P. W. (1975), *Discrete Multivariate Analysis: Theory and Practice* , MIT Press, Cambridge, Massachusetts.

Caron, F. & Mercier, G. (1987), "Rapport d'Opération de la Rivière de la Trinité, 1985", Ministère du Loisir, de la Chasse et de la Pêche, Direction de la Faune Aquatique, Québec.

Chapman, D. G. & Junge, C. O. (1956), "The Estimation of the Size of a Stratified Animal Population", *The Annals of Mathematical Statistics* , 27, 375-89.

Darroch, J. N. (1961), "The Two-Sample Capture-Recapture Census When Tagging and Sampling are Stratified", *Biometrika* , 48, 241-60.

Dempson, J. B. & Stansbury, D. E. (à paraître), "Mark-Recapture Estimate of an Atlantic Salmon, Salmo Salar , Smolt Population in Conne River, Newfoundland, using Partial Counting Fences and a Two-Sample Stratified Technique", Département des Pêches et Océans, St-Jean, T.-N.

Plante, N. (en préparation), "Estimation de la Taille d'une Population Animale à l'Aide d'un Modèle de Capture-Recapture Avec Stratification", Université Laval, Ste-Foy.

Seber, G.A.F. (1982), *The Estimation of Animal Abundance and Related Parameters* (2nd ed.), Charles Griffin & Company Ltd, London & High Wycombe.

Estimation de la variance de l'estimateur par régression du total d'une population finie par la méthode des résidus pondérés

par

Carl-Erik Särndal
Université de Montréal
Montréal, Canada

Bengt Swensson
Université d'Örebro
Örebro, Suède

Jan H. Wretman
Statistique Suède
Stockholm, Suède

Résumé. Cet article porte sur l'estimation de la variance de l'estimateur par regression du total d'une population finie à partir d'un plan de sondage. L'estimateur habituel de la variance calculée par linéarisation de Taylor est exprimé en fonction des résidus de régression pondérés du plan; dans beaucoup d'applications, l'expression ainsi obtenue va à l'encontre de ce qu'un modèle pourrait laisser prévoir. L'estimateur de variance que nous proposons dans cet article associe un autre poids élémentaire, appelé "poids g", à chaque résidu. Ce nouvel estimateur de variance i) donne de bons intervalles de confiance selon un plan, ii) est presque sans biais suivant un modèle de régression acceptable (voir section 5) et iii) se prête bien à l'induction conditionnelle. Nous donnons aussi des exemples de la méthode.

Mots-clés : Induction conditionnelle, intervalles de confiance, induction fondée sur un plan, estimateur par régression, échantillon d'enquête, estimation de la variance.

1. Introduction.

De nombreux ouvrages ont été publiés sur l'estimation de la variance dans les enquêtes par sondage; Wolter (1985) en dresse la nomenclature dans un ouvrage récent. Le fait d'avoir insisté sur les modèles statistiques linéaires a fait avancer la recherche relativement aux deux méthodes d'estimation de la variance dans les enquêtes par sondage : l'estimation par modèle et l'estimation par plan (théorie de la randomisation). Holt et Smith (1979), Royall et Cumberland (1978,1981a,1981b) et d'autres adeptes de l'estimation par modèle ont fait d'importantes observations sur l'estimation de la variance. Certains d'entre eux ont fait ressortir les lacunes de la méthode d'estimation classique. Même pour des estimateurs très courants comme l'estimateur par quotient et l'estimateur par régression simple, la question de l'estimation de la variance demeure insoluble. Les études de Wu (1982), de Wu et Deng (1983), de Deng et Wu (1987) montrent qu'il est difficile de trouver le "meilleur" estimateur de la variance; tout dépend du critère d'efficacité que l'on utilise.

Cet article propose une nouvelle méthode pour estimer la variance de l'estimateur par régression du total d'une population finie. Cet estimateur est utilisé avec des informations supplémentaires concernant une ou plusieurs variables. Avant d'entreprendre cette étude, nous nous sommes posé la question suivante: comment définir un estimateur de la variance de

l'estimateur par régression, qui soit à la fois simple et général, qui donne de bons intervalles de confiance selon un plan et qui soit, du même coup, essentiellement non biaisé par rapport à un modèle de régression hypothètique? La méthode que nous proposons rejoint ces objectifs. La section 5 renferme d'importants résultats concernant les propriétés du nouvel estimateur de la variance suivant un modèle de régression acceptable.

Soit $U = \{1, \ldots, N\}$ une population finie et y_k la valeur de la variable étudiée y pour la kième unité de population; $k = 1, \ldots, N$. Nous voulons estimer le total de population $t = y_1 + \ldots + y_N$. Si A est un ensemble d'unités de population ($A \subseteq U$), nous écrirons $\Sigma_A y_k$ au lieu de $\Sigma_{k\in A} y_k$ (par exemple, $t = \Sigma_U y_k$).

Le premier exemple montre qu'une application inconséquente de la méthode classique d'estimation de la variance peut être discutée et que la méthode que nous proposons dans cet article peut aider à résoudre le problème.

Exemple 1.1. Un échantillon aléatoire simple de taille n est prélevé dans U, puis stratifié. L'estimateur usuel du total de population t est:

$$\hat{t} = \sum_{h=1}^{H} N_h \bar{y}_{s_h} , \qquad (1.1)$$

où $\bar{y}_{S_h} = \Sigma_{S_h} y_k / n_h$ est la moyenne de l'ensemble s_h formé des n_h unités qui se trouvent incluses dans le hième groupe de population, U_h, de taille connue, N_h. Ici, $N = N_1 + \ldots + N_H$; $n = n_1 + \ldots + n_H$. Dans les ouvrages classiques sur l'échantillonnage, par exemple Cochran (1977, p.135), la variance de (1.1), est définie (selon une approximation du premier degré)

$$V_u(\hat{t}) \approx N^2((1-f)/n) \sum_{h=1}^{H} W_h S_{U_h}^2 \quad , \qquad (1.2)$$

où l'indice u signifie "inconditionel" (unconditional) et

$f = n/N$; $W_h = N_h/N$; $S_{Uh}^2 = \Sigma_{Uh}(y_k - \overline{y}_{Uh})^2/(N_h-1)$; $\overline{y}_{Uh} = \Sigma_{Uh}\, y_k/N_h$.

La façon courante de construire un estimateur de variance est d'utiliser la variance comme point de départ. En remplaçant S_{Uh}^2 dans (1.2) par l'équivalent pour un échantillon $S_{Sh}^2 = \Sigma_{Sh}(y_k - \overline{y}_{Sh})^2/(n_h-1)$, nous obtenons la formule souvent proposée

$$\hat{V}_u(\hat{t}) = N^2((1-f)/n) \sum_{h=1}^{H} W_h S_{s_h}^2 \quad , \qquad (1.3)$$

dont la valeur moyenne, pour des échantillons aléatoires simples répétés avec $n_h \geq 2$, est égale à (1.2). Ainsi, l'estimateur (1.3) est "correct" au sens des plans de sondage et beaucoup de statisticiens s'en satisferont. En revanche d'autres comme Holt et Smith (1979),prétendront que l'induction doit dépendre de la configuration effective de l'échantillon $\mathbf{n} = (n_1, \ldots, n_H)'$. Ils critiquent donc l'expression (1.3) parce que, selon celle-ci, la contribution du groupe h à la variance estimée $\hat{V}_u(\hat{t})$ est égale au produit de S_{Sh}^2 par un poids qui n'a aucun rapport avec n_h , soit $N^2W_h(1-f)/n$. Pour appuyer leur argumentation, ce groupe de statisticiens supposent la situation suivante: le groupe h compte un nombre d'observations anormalement faible et la taille effective n_h est de beaucoup inférieure à l'espérance mathématique $E(n_h) = n\, W_h$. Il faudrait donc que la contribution de la strate sous-représentée à la variance estimée soit supérieure à la moyenne. Or, l'expression (1.3) ne permet pas de réaliser cela alors que

$$\hat{V}(\hat{t}) = (1 - f) \sum_{h=1}^{H} N_h^2 S_{s_h}^2 / n_h \quad , \qquad (1.4)$$

par exemple, le permet. Dans l'équation ci-dessus, qui représente l'estimateur de la variance calculé dans l'exemple 4.1 à l'aide de la méthode proposé dans cet article, la variance de strate de l'échantillon S_{Sh}^2 est affectée du poids $(1-f)\, N_h{}^2/n_h$, dont la propension à diminuer lorsque n_h augmente est tout à fait logique au sens de l'approche "conditionnelle".Par ailleurs, les adeptes de l'induction fondée sur un plan peuvent accepter facilement l'expression (1.4) puisque celle-ci est aussi juste que (1.3) en ce sens que l'une et l'autre ont une valeur moyenne au moins approximativement égale à (1.2) pour des échantillons aléatoires simples répétés. L'argument fondé sur la notion de "condition" est important; nous préconisons (1.4) plutôt que (1.3). Passons maintenant à de la théorie générale. □

2. Estimateur par régression générale.

Un échantillon probabiliste s est tiré de U avec des probabilitésde sélection positives π_k = pr(k ∈ s) and $\pi_{k\ell}$ = pr(k and ℓ ∈ s) . Notons que $\pi_{kk} = \pi_k$ pour tous les k. La taille de s , n_s , peut être aléatoire. Cet article porte sur l'estimation de t à partir d'un plan; l'espérance mathématique, le biais et la variance liés au plan deviennent dès lors des valeurs à considérer. On choisit plus souvent un plan d'échantillonnage pour des raisons de commodité que pour des raisons d'efficience (??). Les plans avec des probabilités de sélection inégales ne sont pas rares. Il faut des poids d'échantillonnage $1/\pi_k$, comme dans l'estimateur par régression approximativement non biaisé de t,

$$\hat{t} = \Sigma_s y_k/\pi_k + (\Sigma_U x_k - \Sigma_s x_k/\pi_k)'\hat{B} , \qquad (2.1)$$

où $\hat{B}$ est un q-vecteur de coefficients de régression estimés et $\Sigma_U x_k$ est le total connu des q-vecteurs auxiliaires $x_1, \ldots, x_N$. Cassel, Särndal et Wretman (1976, 1977), Särndal (1980, 1981, 1982), Isaki & Fuller (1982), Wright (1983) et d'autres encore ont étudié des estimateurs de ce genre mais avant même qu'ils tentent d'uniformiser l'expression, diverses formes particulières de (2.1) étaient largement en usage, notamment (1.1), (3.6) et (3.7). Par ailleurs, (2.1) implique une régression sous-jacente

$$y_k = x_k'\beta + \varepsilon_k , \qquad (2.2)$$

pour laquelle il existe une seule hypothèse, à savoir que y est bien expliqué par le vecteur x, de sorte que la variance résiduelle est faible par rapport à la variance totale. Nous ne supposons pas que (2.2) soit le "vrai modèle de superpopulation" qui a donné les valeurs $y_1, \ldots, y_N$. De la même façon, on peut exprimer (2.1) par

$$\hat{t} = \Sigma_U \hat{y}_k + \Sigma_s e_{ks}/\pi_k , \qquad (2.3)$$

où $\hat{y}_k = x_k'\hat{B}$ est la valeur prévue pour la k-ième unité, obtenue en ajustant le modèle de régression décrit plus bas, et le k-ième résidu de la régression est

$$e_{ks} = y_k - \hat{y}_k = y_k - x_k'\hat{B} . \qquad (2.4)$$

Pour une analyse de diverses méthodes de pondération lorsque $\hat{B}$ est calculé, voir Wright (1983), Särndal et Wright (1984). Nous définissons $\hat{B}$ comme l'estimateur ordinaire pondéré selon l'échantillon de

$$B = (\Sigma_U x_k x_k'/c_k)^{-1} \Sigma_U x_k y_k/c_k . \qquad (2.5)$$

De toute évidence B est le résultat d'un ajustement hypothétique par les moindres carrés pondérés de (2.2) à l'ensemble de la population finie (un "ajustement aux données de recensement"); le poids c_k^{-1} est jugé approprié dans un ajustement de ce genre et n'a aucun rapport avec le poids d'échantillonnage π_k^{-1}. Un cas particulièrement simple est celui où $c_k = 1$ pour tous k; cependant, il est parfois justifié d'avoir des valeurs variées pour c_k ; voir exemple 3.1. Comme il est question ici de plan de sondage, B est estimé par l'estimateur équivalent pour un échantillon

$$\hat{B} = (\Sigma_s x_k x_k'/(c_k\pi_k))^{-1} \Sigma_s x_k y_k/(c_k\pi_k). \qquad (2.6)$$

Or, $\hat{B}$ est une fonction de plusieurs sommes d'échantillon du type $\Sigma_s x_{ik}x_{jk}/(c_k\pi_k)$ et $\Sigma_s x_{ik}y_k/(c_k\pi_k)$ où x_{ik} est la valeur de x_i pour la k-ième unité. Chacune de ces sommes est non biaisée selon le plan et, suivant que les quatre premiers moments et les probabilités de sélection répondent à certaines conditions, convergente selon le plan pour la somme appliquée à la population. Par exemple, $\Sigma_s x_{ik}x_{jk}/(c_k\pi_k)$ est convergente selon le plan pour $\Sigma_U x_{ik}x_{jk}/c_k$. Par conséquent, $\hat{B}$ est convergent pour B. Les conditions, que l'on retrouve notamment dans Isaki et Fuller (1982), seront appelées conditions de régularité. Dans le reste de l'article, nous supposons que $\hat{B}$ est exprimé par (2.6) dans la formule de lestimateur (2.3).

3. Les poids g

Nous pouvons maintenant exprimer (2.1) et (2.3) sous la forme suivante

$$\hat{t} = \Sigma_s g_{ks} y_k/\pi_k , \qquad (3.1)$$

à la condition que nous définissions le q-vecteur

$$A_s = (\Sigma_s x_k x_k'/(c_k \pi_k))^{-1} (\Sigma_U x_k - \Sigma_s x_k/\pi_k) ,$$

et que nous posions

$$g_{ks} = 1 + A_s' x_k/c_k . \qquad (3.2).$$

Pour n'importe quel échantillon donné $s \subseteq U$, g_{ks} peut conserver la même définition lorsque k n'est pas inclus dans s; cette définition élargie s'avèrera utile ultérieurement. D'après l'équation (3.1), l'unité k est affectée d'un poids d'échantillonnage π_k^{-1} et d'un "poids g", g_{ks}.

Sans restreindre de façon appréciable la porté de cet article, à partir de maintenant nous n'allons étudier que les cas où c_k dans (2.5) et (2.6) est défini

$$c_k = \lambda' x_k \quad , \qquad (3.3)$$

$k = 1, \ldots, N$, pour un q-vecteur λ indépendant de k de telle sorte que $c_k > 0$ pour tous k. Il est toujours possible d'obtenir (3.3) en additionnant c_k et le vecteur x_k. Lorsque l'équation (3.3) est satisfaite, $\Sigma_s e_{ks}/\pi_k = 0$ pour tous s dans (2.3). En supprimant ce terme, on annule l'effet que pourrait avoir sur l'estimateur (2.3) la présence d'un résidu e_{ks} et d'un poids d'échantillonnage $1/\pi_k$ élevés pour une ou plusieurs unités k. Suivant (3.3), l'estimateur (2.3) devient l'estimateur de <u>projection simple</u>,

$$\hat{t} = \Sigma_U \hat{y}_k/\pi_k = \Sigma_s g_{ks} y_k/\pi_k , \qquad (3.4)$$

où les poids g sont définis pour tout $s \subseteq U$ et tout $k \in U$ par

$$g_{ks} = (\Sigma_U x_k)' (\Sigma_s x_k x_k'/(c_k \pi_k))^{-1} x_k/c_k , \qquad (3.5)$$

c_k étant défini comme en (3.3). Les poids g dépendent de l'échantillon et ressemblent vaguement aux éléments de la "matrice d'estimation" définie

en analyse de régression, voir Hoaglin & Welsch (1978).

Exemple 3.1. L'estimateur par quotient ordinaire pondéré par π peut être déterminé à partir de l'équation (3.4) si x_k est une grandeur scalaire et $c_k \propto x_k$:

$$\hat{t} = (\Sigma_U x_k)(\Sigma_s y_k/\pi_k)/(\Sigma_s x_k/\pi_k) . \qquad (3.6)$$

En l'occurrence, le poids g varie selon s mais non selon k:

$$g_{ks} = (\Sigma_U x_k)/(\Sigma_s x_k/\pi_k) . \qquad \square$$

Exemple 3.2. L'estimateur par régression simple pondéré par π peut être déterminé à partir de l'équation (3.4) si $x_k = (1, x_k)'$, et $c_k = c$, une constante, pour tous k:

$$\hat{t} = N(\tilde{y}_s + \hat{B}(\overline{x}_U - \tilde{x}_s)), \qquad (3.7)$$

où

$$\tilde{y}_s = (\Sigma_s y_k/\pi_k)/\hat{N};\ \tilde{x}_s = (\Sigma_s x_k/\pi_k)/\hat{N};\ \hat{N} = \Sigma_s 1/\pi_k ;\ \overline{x}_U = (\Sigma_U x_k)/N;$$

$$\hat{B} = (\Sigma_s (x_k - \tilde{x}_s)(y_k - \tilde{y}_s)/\pi_k)/(\Sigma_s (x_k - \tilde{x}_s)^2/\pi_k) .$$

Par l'équation (3.5), nous déterminons le poids g de l'unité k

$$g_{ks} = N[\hat{N}^{-1} + (\overline{x}_U - \tilde{x}_s)(x_k - \tilde{x}_s)/\{\Sigma_s (x_k - \tilde{x}_s)^2/\pi_k\}] , \qquad (3.8)$$

qui dépend aussi bien de k que de s. $\square$

Nous allons maintenant énoncer trois propriétés des poids g dont nous nous servirons plus loin:

Propriété 1. Les poids g donnent des "estimations parfaites" lorsqu'ils sont appliqués aux valeurs x_k:

$$\Sigma_s g_{ks} x_k/\pi_k = \Sigma_U x_k . \qquad \square \qquad (3.9)$$

Propriété 2. Pour m=1 et m=2 et n'importe quel échantillon s donné,

$$\Sigma_s (g_{ks})^m c_k/\pi_k = \Sigma_U (g_{ks})^{m-1} c_k , \qquad (3.10)$$

où c_k est défini comme en (3.3). Notons que g_{ks} est défini par (3.5) pour tous $k \in U$. □

Propriété 3.Pour n'importe quel valeur de k, g_{ks} est une variable aléatoire (l'élément aléatoire étant s) dont la distribution est définie par le plan, et pour chaque $k \in U$, g_{ks} converge en probabilité de plan vers l'unité suivant les conditions de régularité: si nous posons

$$T_s = \Sigma_s x_k x_k'/(c_k \pi_k), \quad T_U = \Sigma_U x_k x_k'/c_k ,$$

nous avons $T_U T_s^{-1} \longrightarrow I_{qxq}$, la matrice unité qxq, et par conséquent

$$g_{ks} = \lambda' T_U T_s^{-1} x_k/c_k \longrightarrow \lambda' x_k/c_k = 1 ,$$

où $\longrightarrow$ désigne la convergence en probabilité de plan et où l'équation (3.3) a été utilisée. Pour de grands échantillons, g_{ks} pourra donc avoir une valeur approximative de 1. □

4. L'estimateur proposé et ses propriétés selon un plan.

Par cet article, nous cherchons à améliorer la méthode actuellement utilisée pour estimer la variance de l'estimateur (3.1). Pour cela, il nous faut un estimateur $\hat{V}(\hat{t})$ qui possède de bonnes propriétés par rapport au plan d'échantillonnage qui régit le tirage de l'échantillon et par rapport à un modèle de régression acceptable, et qui soit de la forme simple et d'application générale, c'est-à-dire qu'il convienne à n'importe quel plan et à n'importe quel modèle de régression linéaire. Ce sera d'autant mieux si, au surplus, cet estimateur possède des propriétés rationnelles dans un contexte d'induction conditionnelle.

En ce qui concerne la première exigence, $\hat{V}(\hat{t})$ doit, en règle générale, conduire à un intervalle de confiance valide, défini (à un niveau de confiance inconditionnel approximatif de $1-\alpha$) par

$$\hat{t} \pm z_{1-\frac{1}{2}\alpha} (\hat{V}(\hat{t}))^{1/2} ,$$

la constante $z_{1-(1/2)\alpha}$ étant inférieure à la variable aléatoire normale dans une probabilités de $\alpha/2$. Il faut donc que $\hat{V}(\hat{t})$ soit convergent selon le plan pour $V(\hat{t})$. Comme de nombreux estimateurs ont cette propriété, le choix est encore vaste. Nous allons donc nous tourner vers un modèle de régression acceptable qui nous aidera à restreindre le choix d'estimateurs; notre estimateur devra aussi comporter un biais nul ou quasi nul selon ce modèle. Il s'agit là de la seconde condition énoncée ci-dessus; celle-ci est examinée dans la section 5. Désignons par

E_k le "résidu d'ajustement aux données de recensement":

$$E_k = y_k - x_k'B , \tag{4.1}$$

où B est le vecteur de régression (2.5). Étant donné (3.3), $\Sigma_U E_k = 0$. En nous servant aussi de (3.9), nous pouvons exprimer l'erreur de (3.4) par la formule

$$\hat{t} - t = \Sigma_s g_{ks} E_k/\pi_k . \tag{4.2}$$

L'approximation de la variance de (3.1) par linéarisation de Taylor est

$$V(\hat{t}) \approx \Sigma\Sigma_U \Delta_{k\ell} (E_k/\pi_k)(E_\ell/\pi_\ell) , \tag{4.3}$$

où $\Delta_{k\ell} = \pi_{k\ell} - \pi_k\pi_\ell$. On peut obtenir rapidement la formule (4.3) en posans $g_{ks} = 1$ pour tous k dans (4.2), une approximation pour grands échantillons qui est justifiée par la Propriété 3. Habituellement, on transforme (4.3) en un estimateur de la variance en remplaçant E_k , qui est

inconnu, par l'équivalent pour un échantillon, e_{ks} , qui est défini en (2.4):

$$\hat{V}_1 = \hat{V}_1(\hat{t}) = \Sigma\Sigma_s (\Delta_{k\ell}/\pi_{k\ell})(e_{ks}/\pi_k)(e_{\ell s}/\pi_\ell) \qquad (4.4)$$

où $\Sigma\Sigma_s$ est la forme abrégée de $\Sigma_{k\in s}\Sigma_{\ell\in s}$. L'équation (4.4) est utilisée notamment dans Särndal (1981, 1982) et Särndal et Råbäck (1983). On retrouve π le même concept de pondération de résidus dans SUPERCARP et d'autres programmes servant à l'estimation de la variance dans des enquêtes. La version Yates-Grundy de (4.4), qui est indiquée lorsque le plan est de taille déterminée, est

$$\hat{V}_{YG} = -(1/2)\,\Sigma\Sigma_s (\Delta_{k\ell}/\pi_{k\ell})\,((e_{ks}/\pi_k) - (e_{\ell s}/\pi_\ell))^2 . \qquad (4.5)$$

Kott (1989) indique à quelles conditions l'estimateur (4.5) est convergent selon le plan pour (4.3).

Comme nous le verrons dans les exemples 4.1 à 4.3, les expressions (4.4) et (4.5) ne sont pas les formules idéales lorsque entre en jeu un modèle; cependant, par de légères modifications, on peut améliorer sensiblement l'expression (4.4) sans trop compliquer les calculs:

$$\hat{V}_g = \hat{V}_g(\hat{t}) = \Sigma\Sigma_s (\Delta_{k\ell}/\pi_{k\ell})(g_{ks}e_{ks}/\pi_k)(g_{\ell s}e_{\ell s}/\pi_\ell) \qquad (4.6)$$

Dans l'équation ci-dessus, le résidu est pondéré par π à cause du plan et pondéré par g à cause du modèle; la présence de poids g dans l'expression est tout à fait naturelle étant donné l'équation (4.2). Cette méthode a été proposée dans Särndal (1982). Une version Yates-Grundy de (4.6) peut être utilisée lorsque la taille s est fixe. Comparons les équations (4.6) et (4.4). En vertu de la Propriété 3, g_{ks} converge en probabilité de plan vers l'unité pour tous k. Suivant des conditions additionnelles appropriées, il est possible de montrer que l'écart absolu entre (4.4) et (4.6) converge en probabilité de plan vers 0 de telle sorte que si (4.4) est convergent selon le plan pour (4.3), (4.6) l'est aussi. Par conséquent, le fait de pondérer les

résidus par g ne nuira généralement pas à la convergence selon le plan. Du point de vue d'un plan de sondage, les équations (4.4) et (4.6) sont donc aussi justes l'une que l'autre et la pondération par g peut sembler sans importance. Toutefois, cette opération élimine un inconvénient qui découle souvent de l'équation (4.4), à savoir des expressions contraires à ce à quoi on pourrait s'attendre intuitivement; les exemples 4.1 à 4.3 seront pour nous une occasion d'illustrer ce genre d'inconvénients. Nos tests empiriques indiquent que la formule à poids g (4.6) est au moins aussi efficace que (4.4) pour calculer un taux de couverture empirique proche d'un niveau de confiance nominal de 95% par exemple.

Exemple 4.1. Si nous nous reportons à l'exemple 1.1, posons

$\tilde{y}_{S_h} = (\Sigma_{S_h} y_k/\pi_k)/\hat{N}_h$;

$\hat{N}_h = \Sigma_{S_h} 1/\pi_k$, et considérons l'estimateur de stratification a posteriori "à plan général"

$$\hat{t} = \sum_{h=1}^{H} N_h \tilde{y}_{s_h} \quad . \tag{4.7}$$

Cet estimateur est obtenu à partir de (3.4) si x_k est un H-vecteur qui comprend H-1 éléments de valeur "zéro" et un élément de valeur "unité" qui désigne le groupe de population auquel appartient k, et si $c_k = C_h$, une constante, pour toutes les unités k incluses dans le h-ième groupe. Les poids g requis dans les circonstances pour l'équation (4.6) sont $g_{ks} = N_h/\hat{N}_h$ pour tous $k \in s_h$.

Considérons, comme dans l'exemple 1.1, un échantillonnage aléatoire simple (n unités tirées parmi N; $f = n/N$), et posons

$A_h = (n_h - 1)/n_h)(n/(n-1))$, où n_h est la taille (aléatoire) de s_h. Alors par (4.6) nous avons

$$\hat{V}_g = (1-f)\sum_{h=1}^{H} A_h N_h^2 S_{s_h}^2 / n_h \approx (1-f)\sum_{h=1}^{H} N_h^2 S_{s_h}^2 / n_h \ , \qquad (4.8)$$

où S_{Sh}^2 est défini comme dans l'exemple 1.1, où nous avons relevé l'intérêt du facteur n_h^{-1} au point de vue de l'inférence conditionnelle. En revanche, dans la formule de l'estimateur obtenu à partir de l'équation (4.4), le poids attribué à S_{Sh}^2 est proportionnel à n_h. Ainsi, l'équation produit un résultat contraire à ce à quoi on pourrait s'attendre intuitivement au point de vue conditionnel mais demeure juste au point de vue inconditionnel; l'expression (4.8) a le mérite d'être juste sur les deux plans. □

Exemple 4.2. En ce qui a trait au plan d'échantillonnage de Bernoulli, le fait qu'une unité k soit incluse ou non dans l'échantillon s est déterminé par une expérience de Bernoulli: $\pi_k = Pr(k \in s) = \pi$; $Pr(k \notin s) = 1 - \pi$ pour tous k, les expériences étant indépendantes. Si nous posons $x_k = 1$ et $1/c_k = 1$ pour tous k, l'expression (3.4) donne la moyenne de l'échantillon étendue à l'ensemble,

$$\hat{t} = N \Sigma_s y_k / n_s \ ,$$

par conséquent, $g_{ks} = N\pi/n_s$, où n_s est distribuée selon une loi binomiale. D'après (4.6),

$$\hat{V}_g = ((n_s-1)/n_s)\, N^2(1-\pi)\, S_s^2/n_s \qquad (4.9)$$

$$\approx N^2(1-\pi)\, S_s^2/n_s \ ,$$

où $S_s^2 = \Sigma_s (y_k - \bar{y}_s)^2/(n_s-1)$ est pondéré par $N^2(1-\pi)/n_s$. Ce poids diminue lorsque n_s augmente, ce qui est tout à fait logique au point de vue conditionnel. En revanche, l'équation (4.4) donne dans ce cas

$$\hat{V}_1 = N^2 (1-\pi)(n_s-1) S_s^2/(E(n_s))^2 \ , \qquad (4.10)$$

où $E(n_s) = N\pi$. Le poids de S_s^2 augmente avec n_s, ce qui est difficilement acceptable. Par ailleurs, le "raisonnement classique" peut aboutir à l'expression

$$\hat{V} = N^2 (1-\pi) S_s^2/(E(n_s)) , \qquad (4.11)$$

qui est tout aussi inadéquate puisqu'elle n'a aucun rapport avec n_s . Dans cet exemple, les estimateurs (4.9) à (4.11) sont convergents selon le plan mais seul (4.9) a de l'intérêt au point de vue conditionnel, étant donné n_s. □

Exemple 4.3. Revenons à notre estimateur par quotient ordinaire de l'exemple 3.1 (équation 3.6). L'estimateur de variance à résidus pondérés (équation 4.6) devient

$$\hat{V}_g = ((\Sigma_U x_k)/(\Sigma_s x_k/\pi_k))^2 \; \Sigma\Sigma_s(\Delta_{k\ell}/\pi_{k\ell})(e_{ks}/\pi_k)(e_{\ell s}/\pi_\ell) ,$$

où $e_{ks} = y_k - bx_k)$, $b = (\Sigma_s y_k/\pi_k)/(\Sigma_s x_k/\pi_k)$. Pour un échantillonnage aléatoire simple, en posant $\overline{x}_s = \Sigma_s x_k/n$; $\overline{y}_s = \Sigma_s x_k/n$ et

$$S_{es}^2 = \Sigma_s (y_k - (\overline{y}_s/\overline{x}_s)x_k)^2/(n-1) ,$$

nous obtenons par (4.6) l'estimateur suivant

$$\hat{V}_g = (\overline{x}_U/\overline{x}_s)^2 N^2\{(1 - f)/n\} S_{es}^2 , \qquad (4.12)$$

qui a été analysé notamment dans Wu (1982) et Wu et Deng (1983). Royall et Cumberland (1981a) obtiennent un estimateur à peine différent. C'est pourquoi (4.12) est désormais jugée supérieure à ce qu'on a appelé la "formule classique" (par exemple Cochran (1977, p.155)),

$$\hat{V}_1 = N^2\{(1 - f)/n\} S_{es}^2 ,$$

que l'on obtient par l'équation (4.4) □

Exemple 4.4. Les poids g de l'estimateur par régression simple (3.4)

sont définis en (3.5). Dans le cas précis de l'échantillonnage aléatoire simple, l'estimateur de variance (4.6) devient simplement

$$\hat{V}_g = N^2 ((1 - f)/n) \Sigma_s (g_{ks} e_{ks})^2/(n-1) ,$$

où $e_{ks} = y_k - \bar{y}_s - \hat{B}(x_k - \bar{x}_s)$; $g_{ks} = 1 + n(\bar{x}_U - \bar{x}_s)(x_k - \bar{x}_s)/T_{xx}$, où $T_{xx} = \Sigma_s (x_k - \bar{x}_s)^2$; $\hat{B} = \Sigma_s (y_k - \bar{y}_s)(x_k - \bar{x}_s)/T_{xx}$. Cet estimateur ressemble étroitement (en dépit d'une structure moins complexe) aux estimateurs de variance qu'ont proposés Royall et Cumberland (1978, 1981b) dans leur étude à modèles sur l'estimation robuste de la variance, où il n'est pas question de poids π. □

5. Propriétés de l'estimateur proposé selon un modèle

Nous allons maintenant examiner les propriétés du nouvel estimateur de variance (4.6) en supposant pour $y_1 , \ldots , y_N$ un modèle de régression ξ défini par l'équation

$$y_k = x_k' \beta + \varepsilon_k ,$$

où les ε_k sont indépendants selon le modèle, et de telle sorte que

$$E_\xi(\varepsilon_k) = 0; \quad V_\xi(\varepsilon_k) = \sigma_k^2 = \sigma^2 c_k = \sigma^2 \lambda' x_k \qquad (5.1)$$

pour un vecteur quelconque λ. La structure que nous avons supposée pour la variance $\sigma_k^2 \propto c_k$ justifie la présence de poids $1/c_k$ dans l'équation (2.5). Désignons la moyenne et la variance selon le modèle par E_ξ et V_ξ respectivement. Nous considérons l'estimateur

$$\hat{V}^* = \Sigma\Sigma_s (\Delta_{k\ell}/\pi_{k\ell})(g_k \varepsilon_k/\pi_k)(g_{\ell s} \varepsilon_\ell/\pi_\ell) . \qquad (5.2)$$

Comme les ε_k sont inobservables, $\hat{V}^*$ n'est pas en soi un estimateur de la variance mais il est le prototype de l'estimateur véritable $\hat{V}_g$ défini en

(4.6), où le résidu calculé $e_{ks} = y_k - x_k'\hat{B}$ remplace ε_k. Nous avons

$$E_\xi(e_{ks} - \varepsilon_k) = 0 \; ; \; V_\xi(e_{ks} - \varepsilon_k) = x_k' \, V_\xi(\hat{B}) \, x_k \; .$$

Dans des conditions générales, l'écart $\hat{V}^* - \hat{V}_g$ converge en probabilité de modèle vers zéro.

Dans des études à modèles comme celles de Royall et Eberhardt (1975) et Royall et Cumberland (1978), on cherche d'abord à vérifier si un estimateur de la variance est non biaisé pour l'erreur quadratique moyenne de modèle, $EQM_\xi(\hat{t}) = E_\xi((\hat{t}-t)^2)$.

Nous allons examiner la question dans le contexte de notre article:

Propriétés de $\hat{V}^*$ selon un modèle : Pour n'importe quel échantillon s donné,

(i) $$E_\xi(\hat{V}^*) = \Sigma_s \, (g_{ks} \, \sigma_k/\pi_k)^2 - \Sigma_U \, g_{ks} \, \sigma_k^2 \, , \qquad (5.3)$$

(ii) $$EQM_\xi(\hat{t}) = \Sigma_s \, (g_{ks} \, \sigma_k/\pi_k)^2 - \Sigma_U \, \sigma_k^2 \, , \qquad (5.4)$$

(iii) le biais de modèle relatif (BMR) de $\hat{V}^*$ est

$$BMR_\xi(\hat{V}^*) = [E_\xi(\hat{V}^*) - EQM_\xi(\hat{t})]/EQM_\xi(\hat{t})$$

$$= - \Sigma_U(g_{ks}-1) \, \sigma_k^2/(\Sigma_s \, (g_{ks} \, \sigma_k/\pi_k)^2 - \Sigma_U \, \sigma_k^2). \; \square \qquad (5.5)$$

Notons que pour un échantillon s fixe, la définition de g_{ks} en (3.5) vaut pour tous $k \in U$. Par ailleurs, l'expression (5.5) découle directement de (5.3) et de (5.4) parce que

$$\Sigma_s \, (g_{ks})^m \, \sigma_k^2/\pi_k = \Sigma_U \, (g_{ks})^{m-1} \sigma_k^2 \; , \qquad (5.6)$$

pour $m = 1$ et $m = 2$, cette dernière équation étant dérivée de (3.10) si nous considérons que le modèle (5.1) suppose la relation $\sigma_k^2 \propto c_k = \lambda' x_k$. En outre, $E_\xi(\varepsilon_k \varepsilon_\ell) = 0$ pour tous $k \neq \ell$, de sorte que

$$E_\xi(\hat{V}^*) = \Sigma_s (g_{ks}\,\sigma_k/\pi_k)^2 - \Sigma_s (g_{ks}\,\sigma_k)^2/\pi_k \,.$$

Utilisons maintenant l'équation (5.6) avec $m = 2$ pour transformer le terme négatif de l'expression ci-dessus; nous obtenons alors l'équation (5.3). Pour démontrer (5.4), nous utilisons (3.9) pour en arriver à

$$\hat{t} - t = \Sigma_s (\pi_k^{-1} g_{ks} - 1)\varepsilon_k + \Sigma_{U-s}\, \varepsilon_k \,.$$

Or, une évaluation simple de $E_\xi((\hat{t} - t)^2) = MSE_\xi(\hat{t})$ nous amène à l'équation (5.4) si nous utilisons (5.6) avec $m = 1$.

Les expressions (5.3) à (5.5) nous amènent à faire les commentaires suivants:

(1) Le biais de modèle relatif $BMR_\xi(\hat{V}^*)$ est faible puisque le premier terme du membre de droite de l'équation (5.3) se retrouve aussi, à la même place, dans l'équation (5.4) et qu'il contient le paramètre π_k^{-2} ce qui le rend supérieur au second terme de l'expression, qui est négatif. Par exemple, suivant un échantillonnage aléatoire simple, $BMR_\xi(\hat{V}^*)$ est négligeable si la fraction de sondage $f = n/N$ est elle-même négligeable (voir exemple 5.1 ci-dessous).

(2) $BMR_\xi(\hat{V}^*)$ peut être exactement égal à zéro. Cela arrive dans un cas particulier de l'exemple 4.1, où $g_{ks} = N_h/\hat{N}_h$ pour tous les k inclus dans le groupe h: s'il s'agit d'un échantillonnage aléatoire simple stratifié avec une fraction de sondage fixe n_h/N_h dans la strate h, alors $g_{ks} = 1$ pour tous k et par conséquent $BMR_\xi(\hat{V}^*) = 0$. Dans le cas d'un échantillonnage aléatoire simple avec stratification a posteriori, $BMR_\xi(\hat{V}^*)$ n'est pas nul mais presque.

(3) Lorsque $BMR_\xi(\hat{V}^*)$ n'est pas égal à zéro, il est facile d'ajuster $\hat{V}^*$

de manière à éliminer le biais. Cependant, cette opération ne semble pas vraiment utile dans la pratique car a) elle a habituellement peu d'effet sur les chiffres, b) elle peut compliquer l'expression de l'estimateur de la variance et c) elle suppose un modèle, ce qui est une hypothèse imparfaite. Les propriétés de $\hat{V}^*$ sont suffisamment valables et $BMR_\xi(\hat{V}^*)$ est suffisamment proche de zéro pour que l'on ne se préoccupe pas du faible biais de modèle. Le modèle sert de référence et non de norme pour l'induction fondée sur des plans.

Nous illustrons nos commentaires par l'exemple suivant.

Exemple 5.1. Revenons au cas de l'échantillonnage aléatoire simple (n unités parmi N, de sorte que $\pi_k = n/N = f$ pour tous k), avec l'estimateur par quotient

$$\hat{t} = N\bar{x}_U \bar{y}_s/\bar{x}_s \ .$$

Considérons le modèle

$$y_k = \beta x_k + \varepsilon_k \ , \qquad (5.7)$$

où les erreurs sont indépendantes de sorte que $E_\xi(\varepsilon_k) = 0$, $V_\xi(\varepsilon_k) = \sigma^2 x_k$. La structure de la variance satisfait l'équation (3.3). Comme $g_{ks} = \bar{x}_U/\bar{x}_s$, l'équation (5.2) devient

$$\hat{V}^* = (\bar{x}_U/\bar{x}_s)^2 N^2 ((1 - f)/n) s^2_{\varepsilon s} \ , \qquad (5.8)$$

où

$$s^2_{\varepsilon s} = \Sigma_s (\varepsilon_k - \bar{\varepsilon}_s)^2/(n-1); \ \bar{\varepsilon}_s = \Sigma_s \varepsilon_k/n.$$

Par un calcul, on obtient

$$BMR_\xi(\hat{V}^*) = - (f/(1 - f))((\bar{x}_U - \bar{x}_s)/\bar{x}_{U-s}) \ ,$$

où $\bar{x}_{U-s} = (\Sigma_{U-s} x_k)/(N-n)$. Dans le cas d'un échantillonnage fortement non

compensé, $(\overline{x}_U - \overline{x}_s)/\overline{x}_{U-s}$ peut être très différent de zéro mais malgré cela, $BMR_\xi(\hat{V}^*)$ sera presque nul si f est faible. Si nous remplaçons ε_k dans (5.8) par $e_{ks} = y_k - (\overline{y}_s/\overline{x}_s)x_k$, nous obtenons l'estimateur (4.12), qui est préconisé dans les ouvrages récents. Maintenant, supprimons le biais de modèle de l'estimateur $\hat{V}^*$ défini en (5.8). Posons $\hat{\sigma}^2$ comme n'importe quel estimateur non biaisé de σ^2 selon le modèle. Alors,

$$\hat{V}^{**} = \hat{V}^* + (N\,\overline{x}_U(\overline{x}_U - \overline{x}_s)/\overline{x}_s))\ \hat{\sigma}^2 \qquad (5.9)$$

est non biaisé selon le modèle, c'est-à-dire $E_\xi(\hat{V}^{**}) = EQM_\xi(\hat{t})$. Si nous choisissons en particulier $\hat{\sigma}^2 = S^2_{\varepsilon s}/\overline{x}_s$, l'estimateur non biaisé (5.9) devient

$$\hat{V}^{**} = (\overline{x}_U\overline{x}_{U-s}/(\overline{x}_s)^2)\ N^2\ ((1 - f)/n)\ S^2_{\varepsilon s}\ .$$

La plupart des statisticiens utiliseront indifféremment $\hat{V}^{**}$ et $\hat{V}^*$. Que l'on ait $(\overline{x}_U/\overline{x}_s)^2$ ou $(\overline{x}_U\overline{x}_{U-s})/(\overline{x}_s)^2$ dans l'expression de l'estimateur ne fait pas vraiment de différence. Cependant, la présence de l'un ou l'autre de ces facteurs dans l'espression sera jugée importante car elle permet de hausser l'estimation de la variance lorsque, pour un échantillon, $\overline{x}_s$ est faible par rapport à $\overline{x}_U$. □

Rappelons-nous que l'estimateur $\hat{V}^*$ défini en (5.2) ne pouvait être utile à notre analyse que si nous remplacions ε_k par le résidu calculé $e_{ks} = y_k - x_k'\hat{B}$. Cette substitution a donné l'estimateur proposé (4.6) et a eu effet d'introduire un autre biais de faible importance. Là encore, il est possible de supprimer ce biais mais cette opération a peu d'intérêt en pratique puisqu'elle a habituellement peu d'effet sur les chiffres et que le

modèle sert principalement à guider le choix d'un estimateur de la variance.

En conclusion, nous constatons que Kott (1989) propose pour (2.1) un estimateur de variance qui est a) convergent selon le plan et b) non biaisé pour $E_\xi((\hat{t}-t)^2)$ suivant un modèle de régression hypothétique. Il propose plus précisément de multiplier la formule de Yates- Grundy (équation 4.5) par le "ratio d'ajustement"

$$RA(\hat{V}_{YG}) = E_\xi((\hat{t}-t)^2)/E_\xi(\hat{V}_{YG}) .$$

Cependant, ce ratio peut représenter une expression complexe, même pour des modèles simples, et cette complexité est transmise à l'estimateur de la variance.

REFERENCES

Cassel, C.M., Särndal, C.E. & Wretman, J.H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. Biometrika 63, 615-620.

Cassel, C.M., Särndal, C.E. & Wretman, J.H. (1977). Foundations of Inference in Survey Sampling. New York: Wiley.

Cochran, W. G. (1977). Sampling Techniques, 3rd edition. New York: Wiley.

Deng, L. Y. and Wu, C. F. J. (1987). Estimation of variance of the regression estimator. J. Am. Statist. Assoc. 82, 568-576.

Hoaglin, D. C. & Welsch, R. E. (1978). The hat matrix in regression and ANOVA. The American Statistician 32, 17-22.

Holt, D. & Smith, T. M. F. (1979). Post Stratification. J. R. Statist. Soc. A 142, 33-46.

Isaki, C. T. & Fuller, W. A. (1982). Survey design under the regression superpopulation model. J. Am. Statist. Assoc. 77, 89-96.

Kott, P. S. (1989). Estimating the conditional variance of a design consistent regression estimator. J. Statist. Plan. Inf., to appear 1989.

Royall, R. M. & Cumberland, W. G. (1978). Variance estimation in finite population sampling. J. Am. Statist. Assoc. 73, 351-358.

Royall, R. M. & Cumberland, W. G. (1981a). An empirical study of the ratio estimator and estimators of its variance. J. Am. Statist. Assoc. 76, 66-88.

Royall, R. M. & Cumberland, W. G. (1981b). The finite-population linear regression estimator and estimators of its variance - an empirical study. J. Am. Statist. Assoc. 76, 924-930.

Royall, R. M. & Eberhardt, K. R. (1975). Variance estimates for the ratio estimator. Sankhya C, 37, 43-52.

Särndal, C. E. (1980). On π-inverse weighting versus best linear weighting in probability sampling. Biometrika 67, 639-650.

Särndal, C. E. (1981). Frameworks for inference in survey sampling with applications to small area estimation and adjustment for nonresponse. Bull. Int. Stat. Inst. 49:1, 494-513.

Särndal, C. E. (1982). Implications of survey design for generalized regression estimation of linear functions. J. Statist. Plan. Infer. 7, 155-170.

Särndal, C. E. & Råbäck, G. (1983). Variance estimation and unbiasedness for small domain estimators. Statistical Review 1983:5 (Essays in Honor of Tore E. Dalenius).

Särndal, C. E. & Wright, R. L. (1984). Cosmetic form of estimators in survey sampling. Scand. J. Statist. 11, 146-156.

Wolter, K. M. (1985). Introduction to Variance Estimation. New York: Springer-Verlag.

Wright, R.L. (1983). Finite population sampling with multivariate auxiliary information. J. Am. Statist. Assoc. 78, 879-884.

Wu, C. F. (1982). Estimation of variance of the ratio estimator. Biometrika 69, 183-189.

Wu, C. F. J. & Deng, L. Y. (1983). Estimation of variance of the ratio estimator: an empirical study. In Scientific Inference, Data Analysis and Robustness, ed. by G. E. P. Box et al. New York: Academic Press, 245-277.

SESSION SUR LES APPLICATIONS DE LA STATISTIQUE DANS LE DOMAINE DES TRANSPORTS

Évaluation de l'État du Réseau Routier à l'Entretien du Ministère des Transports

Paul Bergeron, statisticien
Ministère des Transports
Direction de la Recherche
Juin 1989

0. RÉSUMÉ

Un outil de gestion évaluant l'état du réseau routier à l'entretien du ministère des Transports du Québec est développé. Ces indicateurs permettront aux gestionnaires responsables de l'entretien des routes de systématiser la phase d'évaluation dans le cycle de gestion. La communication traitera principalement de la méthode développée et de quelques résultats.

1. INTRODUCTION

1.1 Historique

Le ministère des Transports du Québec a comme principaux objectifs en matière d'entretien: d'assurer la sécurité et le confort des usagers et de protéger les investissements sur un réseau routier de 60 000 km de longueur environ. Le vieillissement du réseau et la diminution des budgets alloués pour l'entretien ont amplifié le besoin d'un outil de gestion uniforme permettant d'ajuster la programmation à ces réalités et d'optimiser l'allocation des ressources. Très souvent, la phase d'évaluation ou de rétroaction est négligée lors de la mise sur pied de programmes même si la perception des gestionnaires en place constitue l'équivalent informel et non uniforme de cette phase. C'est à cette lacune que palliera l'évaluation de l'état du réseau routier. Cet outil de gestion servira aussi, à plus long terme, à la révision des normes d'entretien utilisées.

Ce projet émane d'un rapport interne de M. Claude Marquis, qui proposait des indicateurs de performance obtenus par échantillonnage. L'origine de la demande à notre service, consistait, comme plusieurs autres, à la détermination de tailles d'échantillon permettant une "bonne" représentativité. Nous nous sommes associés au projet et nous impliquons activement dans le développement de la méthode et la collecte des données. Nous avons démarré le travail par une expérience pilote dans deux régions administratives du Ministère, soit le Saguenay/Lac-St-Jean et l'Outaouais.

1.2 Particularités du projet

Certaines informations concernant la structure administrative du Ministère et le réseau routier sont nécessaires à la bonne compréhension du projet. La structure administrative des opérations d'entretien du ministère des Transports a trois niveaux (central, région et district): une cinquantaine de districts groupés dans 15 régions couvrant tout le Québec. A chacun de ces niveaux il y a des

responsables de l'entretien potentiellement concerné par le projet. Par une simplification outrancière, les attributions de chaque niveau se distinguent comme suit: au niveau central sont dévolus des mandats de coordination et d'orientation, au niveau régional, de concertation avec le milieu et de soutien aux districts, qui eux, exécutent les opérations de terrain.

Dans le but d'assurer un meilleur contrôle du développement du projet, certaines exclusions ont été faites. Elles ne diminuent que très peu la portée générale du projet. Nous nous limitons aux routes à chaussée flexible (asphalte) et à chaussée granulaire, ce qui représente au delà de 95% des différents types de surface. La principale exclusion concerne les routes ayant une surface en béton de ciment. Elles sont complexes à évaluer et peu répandues, c'est pourquoi elles ne font pas, pour le moment, l'objet de mesures. De même, l'évaluation de la signalisation et des structures a été mise de côté. Toutefois, les concepts utilisés seront applicables à ces domaines si le besoin en est ressenti. Les types de route utilisés sont les autoroutes, les routes principales, les routes régionales et les chemins dits "municipaux". Les trois premières catégories forment le réseau numéroté d'une longueur d'environ 23 500 km. C'est cette partie importante du réseau qui draine la majeure partie du trafic circulant sur le réseau à l'entretien du Ministère.

La contrepartie externe du projet, c'est-à-dire une évaluation de la perception des utilisateurs du réseau routier, s'intègre à la démarche afin d'obtenir la satisfaction de la clientèle. La présentation portera uniquement sur la partie évaluation interne.

2. MÉTHODE

2.1 Échantillonnage

La méthode utilisée repose sur un échantillonnage stratifié non proportionnel. Les strates sont formées de chaque combinaison des "types de route - types de surface" des districts.

L'univers visé est constitué de tous les kilomètres de route du réseau routier à l'entretien du ministère des Transports du Québec. La population échantillonnée correspond à l'univers visé à l'exception des types de surface autres que "chaussée flexible et granulaire" et des bretelles d'autoroute. Ces exclusions sont mineures et faites pour des raisons pratiques. La population échantillonnée est constituée de sections de route, représentant une partition du réseau routier dont les éléments sont relativement homogènes en ce qui a trait au sujet à l'étude (configuration, trafic, topographie, longueur, etc...). La longueur moyenne d'une section de route est d'environ 2,5 km. Pour des

raisons pratiques et de correspondance avec l'univers visé, les sections échantillonnées sont mesurées sur leur premier kilomètre uniquement. Pour les prochaines collectes, n'importe quel kilomètre d'une section donnée sera l'unité mesurée. Chaque kilomètre mesuré a été divisé en cinq parties égales de 200 mètres de longueur. Ces parties constituent techniquement parlant les "unités répondantes".

Les variables mesurées sont les éléments et sont groupés par activité. Cette nomenclature provient des programmes en vigueur au Ministère. La liste complète des éléments est fournie en annexe. Les quatre activités sont les suivantes: "chaussée et accotement", "drainage", "protection et sécurité" et "abords de route". On comprendra, étant donné l'origine de la dénomination, que les éléments touchant par exemple, la sécurité en général se retrouvent également dans les autres activités que "protection et sécurité". Par exemple, l'élément "pente de l'accotement" contribue au facteur de sécurité même s'il est inclus dans l'activité "chaussée et accotement".

En ce qui concerne les mesures, nous utilisons, comme référence, les normes d'intervention d'entretien en vigueur au Ministère. Cela signifie le seuil au-delà duquel il est nécessaire d'intervenir afin de s'acquitter adéquatement des mandats précités en introduction. L'utilité de ces normes est double, elles permettent d'uniformiser la référence concernant les activités d'entretien (niveau "Relatif") et donnent un aperçu des besoins réels d'intervention (niveau "Absolu"). La première utilité est d'une grande importance, elle constitue la pierre angulaire du projet. C'est cette qualité qui permet des comparaisons entre les résultats de différentes régions administratives ou de districts. Pour un élément donné, nous comparons la mesure prise à la norme, si celle-ci répond à la norme, la variable prend la valeur "1", dans le cas contraire, une valeur de "0". Cette méthode nous permet d'établir la proportion de conformité aux normes. Puisque l'objectif est de construire des indicateurs synthèses, la simplicité des mesures compense largement la "perte" d'informations. D'ailleurs, il est d'autant plus simple de former les gens affectés à la collecte, considérant que le projet sera récurrent (le cycle reste à déterminer) et la collecte de courte durée (quelques semaines).

2.2 Pondérations particulières

Outre la pondération dûe au plan d'échantillonnage, des pondérations subjectives correspondant à l'importance relative des activités, selon l'avis des spécialistes en entretien du Ministère, sont intégrées aux calculs. Elles s'établissent comme suit: "chaussée et accotement" 50%, "drainage" 20%, "protection et sécurité" 20% et "abords de route" 10%. Pour le moment, aucune pondération inter éléments n'est intégrée. Nous prévoyons en ajouter lors du raffinement de la méthode.

2.3 Évaluation de la méthode

Le bilan de l'expérience pilote est très positif. La collecte est simple et relativement peu coûteuse. Les résultats produits comblent les besoins exprimés des gestionnaires impliqués. Toutefois, la subjectivité des normes, en ce qui a trait au niveau "absolu" représente un désavantage mineur dont on doit prendre conscience. Afin de faire accepter que l'échantillonnage utilisé est fiable (en terme de reproductibilité) et de faire connaitre le projet, il est prévu une phase intermédiaire d'extension à un nombre limité de régions supplémentaires (trois) aux deux déjà impliquées. Cette participation, comme celle de l'expérience pilote, s'opère sur une base volontaire. Ceci facilite, stratégiquement, la phase de développement de la méthode et d'implantation future à toutes les régions.

3. RÉSULTATS

Les résultats globaux obtenus dans les deux régions participantes correspondent au discours des spécialistes à l'effet que le réseau routier est en mauvais état. Pour le moment, il ne saurait être question de généraliser ces résultats à l'ensemble du réseau. La proportion de conformité aux normes d'intervention d'entretien des deux régions participantes oscille autour de 65%. Les gestionnaires impliqués au projet ont confirmé l'adéquation de leur perception par rapport aux résultats détaillés. Il en est de même pour les estimations ayant des valeurs extrêmes.

La mesure de variabilité associée aux estimations diffusée aux utilisateurs des données est la valeur permettant de construire un intervalle de confiance à 95%. Celle-ci est la seule utile du point de vue des utilisateurs pour fins de comparaisons entre les différentes estimations.

Nous avons pensé à une stratégie de diffusion de résultats faite sur mesure pour chacun des niveaux de gestion. Ces derniers conditionneront le niveau de détails des résultats fournis. Les gestionnaires des districts recevront les résultats le plus détaillé possible les concernant tandis que les régions et le niveau central auront des résultats plus ou moins globaux. De cette façon, on s'assure que chacun recevra l'information adaptée à ses besoins et selon ses responsabilités administratives.

4. CONCLUSION

La démarche entreprise jusqu'à maintenant s'avère concluante; la méthode développée est éprouvée et les résultats semblent représentatifs de la "réalité". La systématisation de la phase d'évaluation de l'état du réseau routier permettra aux gestionnaires concernés de se concentrer sur l'atteinte de résultats mesurables. La poursuite du projet dans cinq régions nous permettra de tester l'expertise développée dans des milieux plus variés et d'assurer une progression au projet.

Des efforts devront être investis quant aux "outputs" produits et à leur forme. Les besoins des trois niveaux de gestionnaires devront être suivi dans le temps afin de s'ajuster aux changements. Il faudra aussi implanter un processus d'assurance de la qualité des données à toutes les étapes et en particulier lors de la collecte. Il est souhaitable de tendre vers des mécanismes d'autorégulation plutôt que vers des méthodes de contrôle à connotation péjorative.

ANNEXE
LISTE DES ACTIVITÉS, SOUS-ACTIVITÉS ET DES ÉLÉMENTS MESURÉS

ACTIVITÉ	SOUS-ACTIVITÉ	ÉLÉMENT
Chaussée et accotement	. Chaussée en matériaux flexibles	Indice de confort
		Rebondissement
		Profil transversal
		Ornières
		Dépressions ou soulèvements sur 3 mètres
		Trous
		Fissuration
	. Accotement asphalté	Trous
		Fissuration
		Pente transversale
	. Accotement en matériaux granulaires	Pente transversale
		Affaissement
		Andains
		Niveau par rapport à la chaussée
	. Chaussée en matériaux granulaires	Épaisseur de la fondation supérieure
		Trous, "nids de poule", "planche à laver"
		Bombements et andains
		Ornières
		Abat-poussière
Drainage	. Fossés	Obstruction
		Profil du fossé par rapport à la surface
	. Conduites et ponceaux	Obstruction
		Dépôts de sédiments
		Bris
	. Regards	Dépôts de sédiments
		Niveau du couvercle
	. Puisards	Dépôts de sédiments
		Niveau de la grille
Protection et sécurité	. Glissières	Manque
		Profil et hauteur
		Alignement
		Emplacement
		Bris ou détérioration
	. Bordures	Hauteur et état général
	. Clôtures	État général
Abords de route	. Broussailles	Hauteur
	. Entrées privées	Géométrie
		Drainage

IDENTIFICATION DE PROFILS TYPES DE CIRCULATION POUR FINS D'ESTIMATION

Martine Désilets et Claude Ouimet
Ministère des Transports du Québec

1. Introduction

Le ministère des Transports est intéressé à connaître les débits de circulation sur tout le réseau routier numéroté du Québec pour ses besoins de planification. Qu'il soit question d'ajouter un tronçon de route ou d'en modifier la configuration, de planifier les besoins d'entretien du réseau ou d'évaluer des taux d'accidents, cela nécessite une connaissance assez précise des débits de circulation et de leurs fluctuations.

La statistique la plus couramment utilisée par les ingénieurs pour exprimer le niveau moyen de la circulation sur une route est le débit journalier moyen annuel, noté D.J.M.A. Pour estimer les D.J.M.A., on procède à deux types de comptages: les comptages permanents et les comptages partiels. Les stations permanentes sont composées d'un système de détection et d'un compteur. Le système de détection repère les véhicules qui circulent dans les deux directions en un point précis de la route et le compteur enregistre cette fréquence à toutes les heures. Pour les stations permanentes, on connaît donc les débits pour chaque heure de l'année alors que les comptages partiels ne sont effectués que sur une courte période (quelques jours), à quelques reprises dans l'année. On estime alors le D.J.M.A. à partir des modèles que nous fournissent les comptages permanents. Notons que le D.J.M.A. évalué à un site de comptage est valide pour toute la section de route avoisinante.

Dans un objectif de rationnalisation du plan de comptage, le ministère a entrepris de réviser l'ensemble du processus de production des débits de circulation, tant du point de vue opérationnel (types de compteurs, détection, cueillette par télémétrie) que de celui des méthodes d'estimation, et ce, afin d'assurer un certain niveau de fiabilité aux données.

2. Démarche méthodologique

Pour réaliser cette étude, on a proposé une démarche méthodologique en quatre étapes: étude de stabilité temporelle des profils de circulation, groupement des stations permanentes d'après leurs profils, lien avec les caractéristiques des sections de route où les stations se situent, plan d'échantillonnage et méthodes d'estimation. Les résultats de la première étape ainsi que le début de la deuxième étape ont été présentés au colloque de l'an dernier. Nous insisterons donc ici sur les étapes réalisées au cours de cette année, i.e. la suite de la deuxième étape et la troisième étape.

2.1 Etude de stabilité temporelle des profils de circulation

Cette étude visait à vérifier l'hypothèse voulant que le profil de circulation d'une station permanente soit stable d'une année à l'autre. La figure 1 illustre différents profils de circulation à l'aide de courbes de variations mensuelles et journalières des débits de circulation moyens par rapport au D.J.M.A. Pour vérifier si on retrouve ces mêmes courbes d'une année à l'autre, on a choisi de considérer les 98 stations permanentes pour lesquelles on a au moins trois années complètes de données sur la période 1981 à 1985. On a calculé des coefficients de variation pour différents niveaux et, à l'aide de critères préétablis, on a déterminé que 86 stations permanentes étaient stables au niveau mensuel; ce qui corroborait notre hypothèse. Au niveau hebdomadaire, après avoir éliminé les semaines contenant des congés fériés, l'analyse nous a révélé que certaines périodes de l'année sont plus stables. Ce résultat sera utile à la détermination des périodes propices aux comptages partiels.

2.2 Groupement des stations permanentes

Ayant vérifié l'hypothèse de stabilité, on peut maintenant procéder au groupement des stations permanentes ayant des profils semblables, afin d'en dégager des profils types de circulation.

L'examen des méthodes utilisées dans d'autres administrations publiques au Canada et aux Etats-Unis s'est avéré décevant car la plupart de ces méthodes, comme celle du Bureau of Public Roads notamment, ne tiennent compte que des variations mensuelles. Etant à la recherche d'une méthode statisti-

que nous permettant de tenir compte des 365 facteurs journaliers (débit journalier/D.J.M.A.), on a pensé utiliser les techniques de classification automatique ("cluster analysis"). Cependant, afin de pouvoir utiliser les statistiques nous permettant de déterminer le nombre idéal de groupes, les variables à classifier doivent être le plus possible indépendantes.

Afin de remédier au problème de dépendance, nous utiliserons une définition des journées de midi à midi pour l'analyse de nos données; l'expérience nous montrant que la circulation suit un comportement assez semblable d'une journée à l'autre, du lundi midi au vendredi midi. Nous obtenons ainsi quatre jours de semaine, qu'on souhaite être assez semblables pour pouvoir les regrouper, et trois jours de fin de semaine très différents des jours de semaine. Afin de réduire notre ensemble de 365 variables à un plus petit ensemble de variables le moins correlées possible, nous allons procéder à des analyses de variance et à des analyses en composantes principales. Ensuite, nous choisirons les variables à retenir pour la classification automatique. Les résultats de ces analyses sont détaillés dans les trois sections suivantes.

2.2.1 Analyse de variance

Dans le but de réduire notre ensemble de données, notre première tentative a été d'utiliser l'analyse de variance. On a ainsi cherché à comparer l'effet des mois et des jours de semaine sur les débits de circulation.

Chaque station permanente étant considérée comme une expérience isolée, on a procédé à une ANOVA par site.

Le modèle proposé en est un non balancé à deux facteurs fixes, l'effet de l'année étant non significatif.

Toutefois, comme les hypothèses de normalité et d'égalité de variances ne sont pas respectées, même suite à des transformations de variables, on va emprunter une autre méthode d'analyse.

2.2.2 Analyse en composantes principales (ACP)

Le deuxième outil utilisé ayant pour but d'éliminer la corrélation présente entre nos variables, c'est l'analyse en composantes principales. Cette méthode nous propose de résumer l'information contenue dans l'ensemble de données. Les nouvel-

les variables créées sont des combinaisons linéaires des variables de départ. On peut ainsi réduire considérablement notre groupe de variables tout en conservant un bon pourcentage de la variabilité présente dans l'ensemble de données. De plus, les variables ainsi créées sont indépendantes par définition.

On procède donc à une ACP différente pour chacune des cinq années à l'étude. Tel que mentionné précédemment, il nous est difficile de travailler avec une année moyenne, vu la non concordance des jours fériés d'une année à l'autre.

Pour les besoins de notre analyse, les variables utilisées sont les 365 facteurs journaliers, soit les rapports des débits journaliers sur le D.J.M.A. Donc, pour les années 1981 à 1985, on procède respectivement avec 73, 89, 91, 83 et 78 stations permanentes.

Si on s'attarde aux résultats maintenant, on peut voir l'apport moyen de chacun des axes, calculé à l'aide des valeurs propres. Les contributions présentées dans le tableau 1 sont des moyennes sur les cinq années à l'étude, les résultats étant comparables à tout le moins pour les trois premiers axes.

Le premier axe contribue pour plus de 43% de la variabilité de l'information présente dans nos données; il faut en conserver cinq pour expliquer plus de 75%.

Tableau 1

Axe	Contribution (%)	Contribution cumulative (%)
1	43,5	43,5
2	18,5	62,0
3	6,2	68,2
4	4,9	73,1
5	3,2	76,3

Les corrélations temporelles des composantes principales sur les cinq années sont supérieures à ,88 pour les deux premiers axes, alors qu'elles varient de ,60 à ,88 pour le troisième axe. Pour ce qui est des corrélations des quatrième et cinquième axes, elles ne sont pas significatives en général. C'est pourquoi on ne conserve que trois axes.

Si on passe maintenant à la signification des axes retenus, on constate que le premier axe oppose les stations permanentes dont les profils comportent peu de variation, tant au niveau mensuel que journalier, à celles dont les profils ont une grande amplitude de variation. Les graphiques de la figure 1 expriment bien ce phénomène. La courbe "urbain" laisse voir une certaine uniformité des débits de circulation selon les mois et les jours de la semaine. La courbe "rurale" quant à elle nous dévoile des flux de circulation beaucoup plus forts l'été que l'hiver. Au niveau journalier, les débits des trois jours de fin de semaine sont nettement supérieurs à ceux du lundi au jeudi.

Au sujet du deuxième axe, on remarque qu'il permet de faire une certaine distinction entre les stations permanentes dont les profils ont une grande amplitude de variation. Parmi celles-ci, cet axe oppose les stations permanentes à variations journalières très fortes à d'autres dont les variations mensuelles sont très élevées. Les graphiques de la figure 2 démontrent bien ce fait. Les routes à profil "récréatif" sont beaucoup plus achalandées du vendredi au dimanche. De leur côté, les routes à profil "touristique" accueillent beaucoup plus de trafic l'été que l'hiver.

Quant au troisième axe, on ne lui trouve pas d'interprétation spéciale, les éléments les plus marquants des vecteurs propres correspondant surtout à des jours de fin de semaine.

Donc, en résumé pour l'analyse en composantes principales, nous ne retenons que les trois premiers axes qui expliquent plus de 68% de la variabilité totale.

2.2.3 Classification automatique

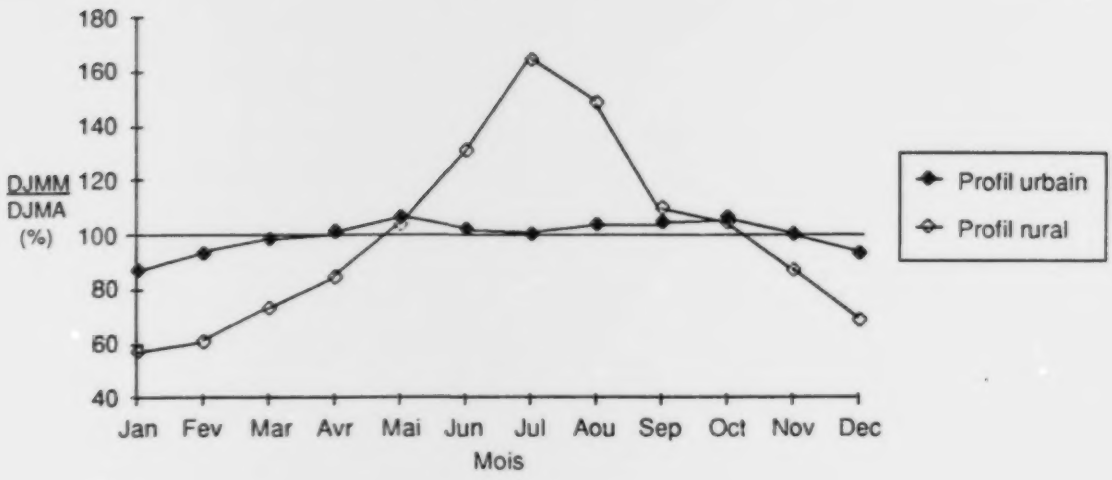
Maintenant qu'on satisfait les conditions d'utilisation de la classification automatique, soit d'avoir un ensemble de variables indépendantes, on peut tenter de regrouper nos stations permanentes selon les trois premières composantes principales des profils de circulation.

La méthode d'analyse utilisée fait intervenir la notion de distance entre les éléments en présence, à savoir les stations permanentes. Notons également que les différentes méthodes développées au fil des années ont leur propre définition de distance.

Pour notre part, nous utilisons une méthode hiérarchi-

Figure 1

Variations mensuelles du débit journalier moyen par rapport au D.J.M.A.



Variations journalières du débit journalier moyen par rapport au D.J.M.A.

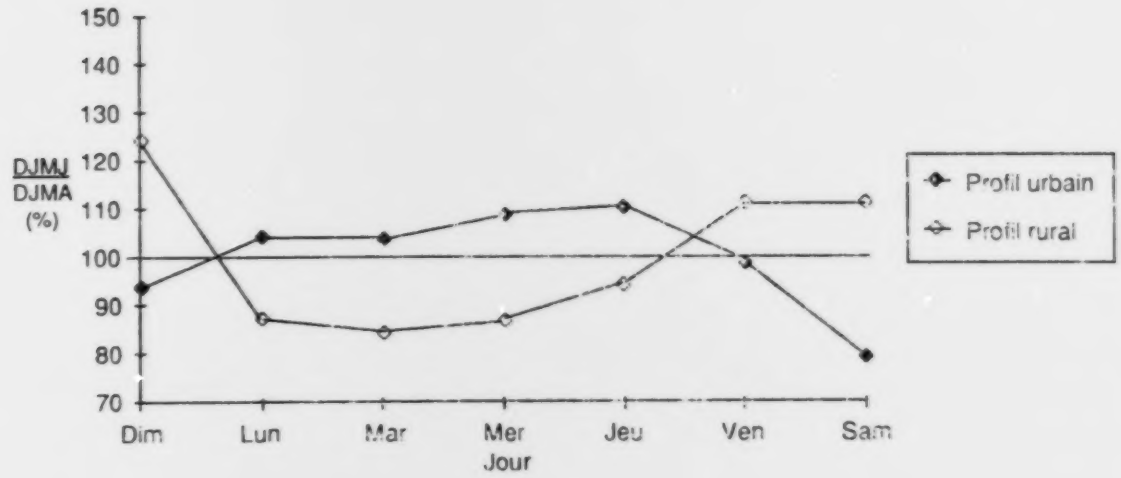
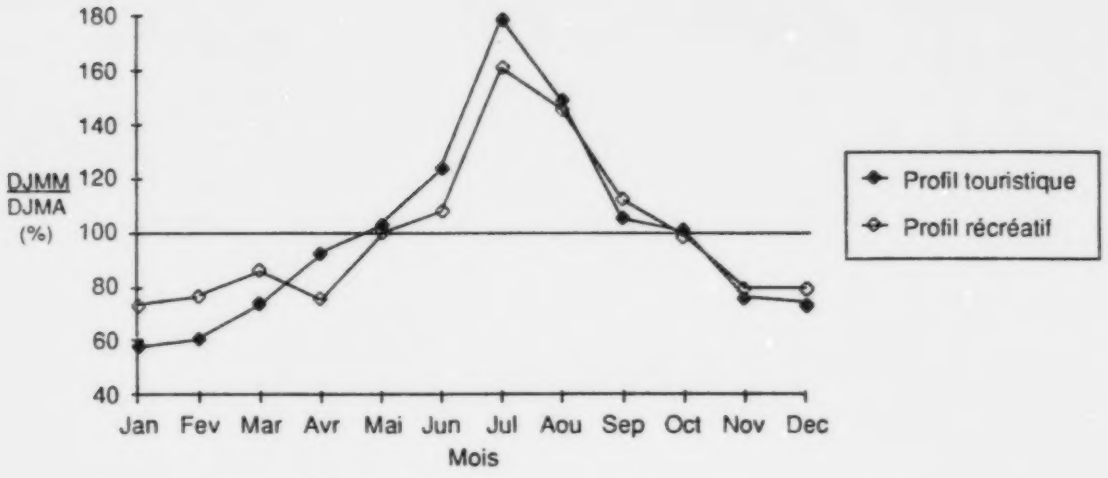
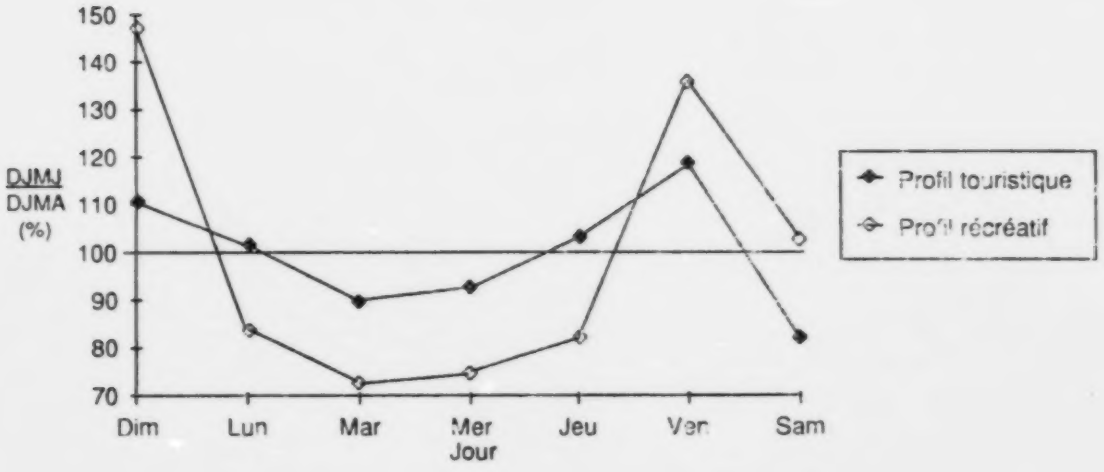


Figure 2

Variations mensuelles du débit journalier moyen par rapport au D.J.M.A.



Variations journalières du débit journalier moyen par rapport au D.J.M.A.



que, i.e. qu'on fait appel à la matrice des distances entre toutes les paires de stations. La méthode est également dite agglomérative, car les étapes de groupement sont consécutives et dépendantes entre elles. On débute donc avec 98 groupes formés d'une station permanente chacun, et le processus d'agglomération se termine quand toutes les stations ne forment plus qu'un seul groupe.

Les deux mesures de distance retenues pour faire nos groupements sont la méthode de Ward et celle du centroïde. La première favorise une variance intragroupe minimale, mais en contrepartie, elle est biaisée en faveur de la formation de groupes avec sensiblement le même nombre d'éléments; elle est également sensible aux valeurs aberrantes. La méthode du centroïde est une variante de celle de Ward, qui utilise la distance euclidienne et qui est plus robuste aux données aberrantes. Nos groupements définitifs sont donc basés sur la combinaison des résultats des deux méthodes.

L'utilisation d'une méthode hiérarchique agglomérative nous permet de conserver plusieurs partitions. On en a obtenu trois, soit une en cinq groupes, une autre en six et une troisième en neuf groupes. On peut voir à la figure 3 comment la partition en cinq groupes se disperse selon les deux premières composantes principales.

2.3 Lien avec les caractéristiques des sections de route

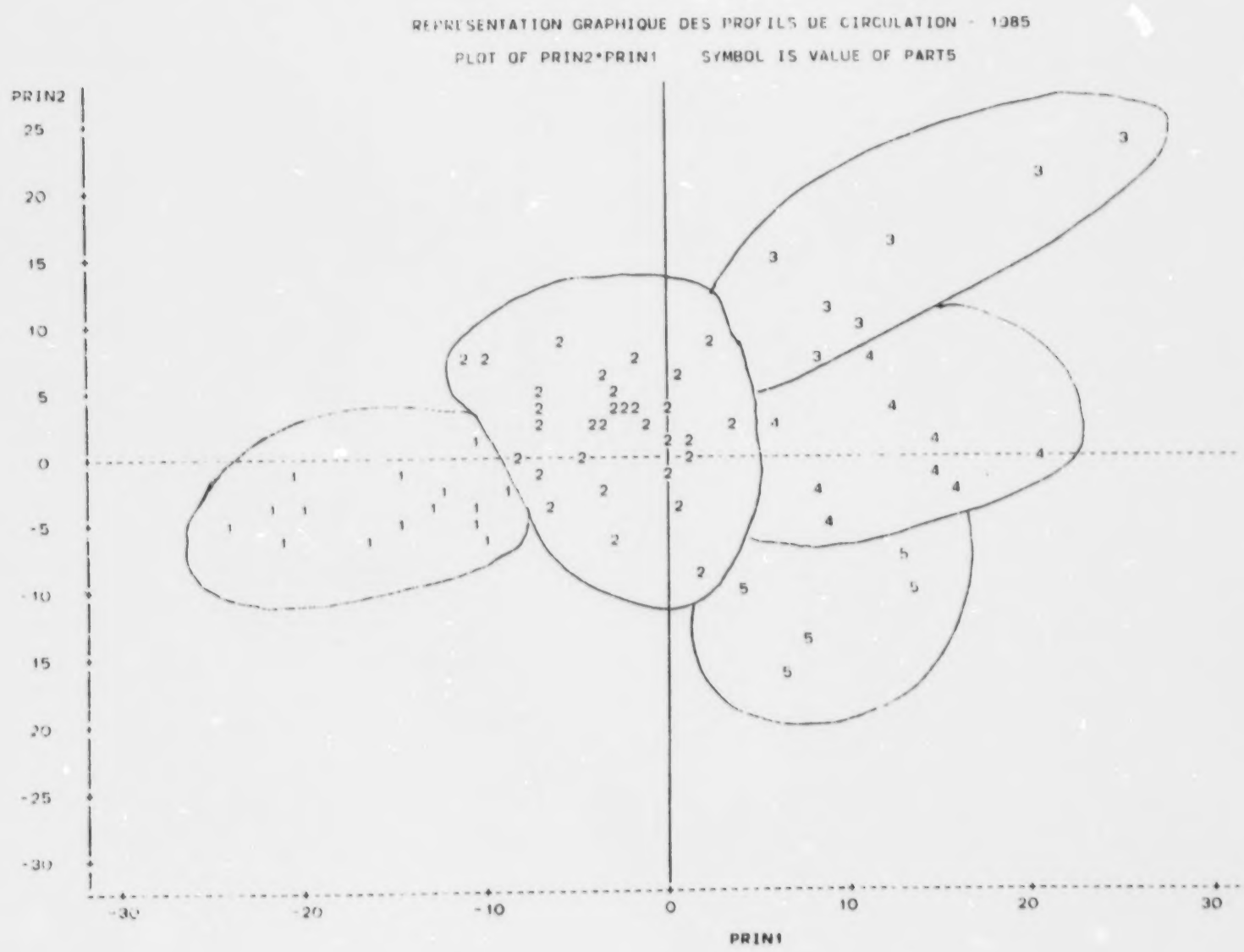
Une fois les groupements obtenus, on cherche à savoir si les stations permanentes d'un même groupe ont des caractéristiques en commun. Cela nous permettrait aisément d'associer une section de route dépourvue de compteur permanent à un profil type. On aurait également un outil pour vérifier s'il existe d'autres profils types de circulation non représentés par les stations en place.

On a procédé à l'inventaire des informations disponibles au ministère et les caractéristiques suivantes nous apparaissent les plus intéressantes.

1) Type de milieu

- urbain
- rural

Figure 3


REPRESENTATION GRAPHIQUE DES PROFILS DE CIRCULATION - 1985
PLOT OF PRIN2*PRIN1 SYMBOL IS VALUE OF PART5
PRIN2
25
20
15
10
5
0
-5
-10
15
20
-25
-30
-30
-20
-10
0
10
20
30
PRIN1

2) Classe de route

- autoroute
- route principale
- route secondaire

3) Classe de D.J.M.A.

La combinaison des trois caractéristiques ne nous permet pas d'établir une concordance avec les profils types identifiés. C'est pourquoi, on pense à long terme se tourner vers la vocation de la route. En effet, une issue intéressante serait de recueillir de l'information sur la vocation intrinsèque de la route, à l'aide d'enquêtes O-D (origine et destination). On pourrait ainsi ériger un système basé sur: l'identification de générateurs de trafic, les types d'usagers, les raisons des déplacements, leur longueur, etc...

3. Conclusion

Après avoir identifié sommairement cinq profils types de circulation, nous passons maintenant à l'étape de développement des méthodes d'estimation des D.J.M.A. pour les comptages partiels, et de détermination du plan d'échantillonnage. Etant donné l'absence d'une ou plusieurs caractéristiques nous permettant de bien distinguer les profils types de circulation, nous allons tenter de trouver une règle d'association des sections de route aux profils types, basée sur des relevés partiels de circulation. Il faudra donc déterminer ce que devront être:

- la période de cueillette
- la fréquence des relevés
- la durée d'un relevé
- la combinaison des jours composant la période de relevés
- l'aménagement des heures de comptage

Nous espérons que cette méthode produira des résultats satisfaisants d'ici à ce que nous ayons obtenu plus d'information pour caractériser les profils de circulation.

SESSION SUR LA CONFIDENTIALITÉ DES DONNÉES ET LE COUPLAGE DE FICHIERS ADMINISTRATIFS

Fichiers publics de microdonnées: le problème de l'identification et de la divulgation et quelques techniques palliatives

Johanne DENIS

Statistique Canada

Mots clés: confidentialité, protection de la vie privée, loi sur la Statistique, unicité

RÉSUMÉ

Statistique Canada fait présentement face à une demande croissante en ce qui a trait à la diffusion de fichiers informatiques de microdonnées. Toutefois, le problème de l'identification et de la divulgation de l'information en ce qui concerne la vie privée des individus en complique de beaucoup la production. Pour pallier ce problème, le bureau doit tenter d'évaluer le "risque" d'identification et de divulgation de chaque fichier de microdonnées qu'il projette de diffuser, et il doit apporter les rectifications jugées nécessaires.

Nous introduisons d'abord le sujet en distinguant des facteurs et méthodes qui favorisent l'identification. À titre d'application, nous présentons le problème dans le cas des fichiers de microdonnées à grande diffusion du recensement de la population. Quelques techniques palliatives sont décrites.

1. INTRODUCTION

Nous utilisons le terme "microdonnée" lorsque nous référons à un fichier dans lequel chaque enregistrement fournit des données relatives à un individu, un ménage, un établissement ou à une toute autre unité. Un fichier informatique contenant les enregistrements provenant d'une enquête ou d'un recensement par exemple, forme un fichier de microdonnées. Habituellement, l'information contenue dans ces fichiers est résumée sous forme de tableaux ou agrégée pour produire les statistiques requises dans les publications et les divers rapports.

Toutefois, aussi utiles que sont ces agrégations de données, les microdonnées sous-jacentes offrent une puissance d'analyse beaucoup plus importante. En effet, certaines techniques d'analyse de données peuvent être appliquées à ces microdonnées permettant l'obtention d'informations particulières aux besoins des utilisateurs. Ainsi, Statistique Canada fait présentement face à une demande croissante en ce qui a trait à la diffusion de tels fichiers informatiques de microdonnées. Toutefois, avant d'autoriser la diffusion du produit, le Bureau doit être assuré que toutes les mesures ont été prises pour assurer la protection face à l'identification des unités enquêtées.

Nous introduisons d'abord les concepts de l'identification et de la divulgation tels que définis par la loi sur la Statistique. Des facteurs et méthodes permettant l'identification sont ensuite décrits. À titre d'application, nous présentons le problème dans le cas du fichier de microdonnées à grande diffusion sur les ménages et les

logements du recensement de la population. Quelques techniques palliatives sont décrites.

2. LOI SUR LA STATISTIQUE ET LE PROBLÈME DE L'IDENTIFICATION ET DE LA DIVULGATION

Au Canada il existe une loi statuant que toute donnée diffusée doit toujours assurer la protection face à l'identification d'un individu, d'une entreprise ou d'une organisation.

Ainsi, selon la loi sur la Statistique, article 16:

"...aucune personne qui a été assermentée en vertu de la loi sur la Statistique ne doit révéler ni sciemment faire révéler, par quelque moyen, des renseignements obtenus en vertu de la présente loi de manière qu'il soit possible, grâce à de telles révélations, de rattacher à un particulier, à une entreprise ou à une organisation identifiables les détails obtenus dans un relevé qui les concerne exclusivement."

De manière à protéger l'identification des unités, les fichiers de microdonnées ne contiennent aucun identificateur direct tels nom, adresse, numéro d'assurance sociale. Toutefois, une combinaison de variables telles la localisation géographique, l'âge, l'origine ethnique, la profession et le revenu, si suffisamment détaillée, peut permettre l'identification d'un individu connu à l'avance par l'investigateur c'est-à-dire, l'établissement d'une relation entre une unité connue et un enregistrement du fichier de microdonnées.

Ainsi, la diffusion de microdonnées implique toujours certains risques d'identification. Une conséquence de l'identification est la divulgation d'informations particulières se rapportant à une unité. On entend ici par information particulière les valeurs de variables qui appartiennent au domaine privé de l'unité, valeurs qu'elle ne désire pas voir révélées.

On peut facilement imaginer les conséquences d'un bris de confidentialité pour le Bureau. Si les répondants ne font plus confiance à l'agence statistique, ils ne répondront plus aux questions d'enquêtes. Les taux de non-réponse iront en augmentant avec les conséquences qu'on imagine.

Ayant statué que l'identification d'une unité est inacceptable, nous pouvons toutefois soulever la question suivante: jusqu'à quel niveau un fichier informatique doit-il être protégé contre l'identification? Le mandat de toute agence statistique étant de recueillir et de diffuser de l'information statistique, nous faisons ici face au dilemme suivant: la liberté de l'information versus le droit à la vie privée; dilemme qui représente vraiment le noyau du problème de la violation de confidentialité.

3. FACTEURS ET MÉTHODES FAVORISANT L'IDENTIFICATION

De manière à protéger un fichier de microdonnées contre l'identification, nous devons isoler les facteurs causals et déterminer

certaines méthodes qui la favorisent. Ainsi, le concept d'unicité est d'une grande importance. Nous dirons qu'un "enquêté" est unique dans la population s'il est le seul à posséder un ensemble particulier de caractéristiques. Nous dirons qu'un "enquêté" est unique dans l'échantillon s'il est le seul dans l'échantillon à posséder cet ensemble particulier de caractéristiques. Ainsi, un "enquêté" unique dans la population sera également unique dans l'échantillon, s'il a été sélectionné. Nous remarquons toutefois que l'unicité dans l'échantillon n'implique pas l'unicité dans la population.

Il est clair qu'un investigateur intéressé par exemple, par les individus uniques dans la population qui ont été sélectionnés dans l'échantillon n'aura qu'à se concentrer sur les enregistrements possédant un ensemble unique de caractéristiques.

L'information géographique contenue dans le fichier et la taille de l'échantillon jouent également un rôle important dans l'identification d'une unité. Ainsi, une petite région géographique combinée à une faible taille d'échantillon entrainera une chance plus élevée d'observer des unicités et par conséquent d'identifier certaines unités.

Ainsi différentes méthodes vont permettre l'identification d'une unité, entre autres:

i) identification par unicité dans la population et sélectionné dans l'échantillon;

ii) identification par appariement:

supposons qu'un investigateur possède un registre contenant les mêmes caractéristiques que dans l'échantillon en plus des noms et adresses. Si le registre offre une énumération complète de la population, il pourrait être possible d'apparier de façon unique un enregistrement échantillonné à un enregistrement du registre. De façon plus large, supposons que deux ou plusieurs fichiers de microdonnées provenant de la même population soient disponibles, que plusieurs des caractéristiques des fichiers soient communes, et qu'en plus chaque fichier offre une série de caractéristiques différentes. Par appariement de caractéristiques clés, il serait donc possible d'apparier un ou plusieurs enregistrements et d'augmenter ainsi la possibilité d'identifier une unité particulière;

iii) identification par connaissance de la réponse:

supposons qu'un investigateur sache qu'un individu particulier a participé à l'enquête et que conséquemment, les données se rapportant à cet individu devraient être dans le fichier, l'identification est très facile si cet individu est unique dans l'échantillon (pas nécessairement dans la population).

4. APPLICATION AUX FICHIERS DE MICRODONNÉES À GRANDE DIFFUSION DU RECENSEMENT

Depuis le recensement de 1971, Statistique Canada offre des échantillons de microdonnées sélectionnés à même les bases de données du recensement.

Au recensement de 1986, 3 fichiers de microdonnées sont proposés soit, un fichier sur les ménages et les logements, un fichier sur les individus et un fichier sur les familles. À ce jour, le fichier de microdonnées sur les ménages et les logements a été créé et diffusé.

En guise d'application, nous présentons donc le problème de l'identification dans le cas du fichier de microdonnées à grande diffusion sur les ménages et les logements et nous décrivons les méthodes qui ont été utilisées pour tenter de protéger l'identification des unités échantillonnées.

4.1 Aperçu méthodologique du fichier de microdonnées à grande diffusion sur les ménages et les logements

Nous faisons d'abord une brève description de la méthodologie d'échantillonnage du fichier.

La population visée par le fichier comprend tous les ménages privés dans les logements occupés au Canada, en excluant les ménages situés sur des réserves indiennes. L'identification géographique est limitée aux provinces et aux régions métropolitaines de recensement (RMR) soit une région urbaine de 100,000 habitants et plus. De plus, un code rural/urbain est disponible pour les régions n'étant pas définies comme RMR ou agglomération de recensement, soit une région urbaine de 10,000 habitants ou plus. L'Ile-du-Prince-Édouard, le Yukon et les Territoires du Nord-Ouest n'offrent toutefois pas cette ventilation. Un critère de population minimale par RMR est imposé; soit celui d'une population minimale de 50,000 ménages. Ainsi, les RMR violant cette condition ne sont pas identifiées dans le fichier.

Une méthode d'échantillonnage à deux degrés a été utilisée. Le premier degré se composait tout simplement de l'échantillon de un cinquième du recensement (échantillon 2B). Au deuxième degré d'échantillonnage, un échantillon de ménages était sélectionné de l'échantillon 2B de façon systématique avec probabilité proportionnelle à une mesure de poids des ménages et ce, à l'intérieur de strates définies par la région géographique (RMR et non RMR).

Le taux de sondage du deuxième degré a d'abord été fixé à 1/100; toutefois, étant donnée la taille réduite de certaines RMR, une taille d'échantillon minimale de 2,500 par RMR a été établie occasionnant l'utilisation de fractions de sondage différentes par RMR. De façon globale, le taux de sondage a été d'environ 1/75 (taille d'échantillon de 115,347 ménages). L'échantillon obtenu était auto-pondéré à l'intérieur des strates.

Du point de vue du contenu, le fichier procure des données à la fois sur le ménage en tant que tel, sur le logement qu'il occupe, ainsi que

de nombreuses informations démographiques, sociales et économiques sur le soutien du ménage, son conjoint et la famille économique dont il est membre.

4.2 Méthodes adoptées pour protéger l'identification des ménages

À Statistique Canada, il existe une politique concernant la diffusion des microdonnées qui stipule entre autres, qu'avant d'autoriser la diffusion d'un fichier de microdonnées, le Bureau doit être convaincu que toutes les mesures qui s'imposent ont été prises pour éviter l'identification des unités.

Selon la description faite précédemment de différentes méthodes permettant l'identification d'une unité, dans le contexte du présent fichier, un effort considérable a été apporté 1) pour contrer l'identification par unicité dans la population et sélectionné dans l'échantillon, et 2) pour tenter de résoudre les cas où la fréquence observée était jugée trop petite. Les procédures suivantes ont donc été adoptées:

i) examiner la distribution univariée de chaque variable et comparer, pour chaque catégorie de cette variable, le compte de la population et le compte de l'échantillon. Cette comparaison est effectuée à l'intérieur de chaque région géographique identifiée sur le fichier;

ii) si une ou plus d'une catégorie de variable possède des comptes relativement bas d'échantillon et de population pour une quantité importante de géographie, alors suggérer un regroupement de catégories;

iii) si des fréquences relativement petites sont observées dans seulement quelques régions géographiques, examiner quelques distributions croisées et juger du potentiel d'identification de la variable. En fait, le regroupement de catégories pour tenir compte de problèmes rencontrés dans seulement quelques régions géographiques n'est pas souhaitable, dans ces cas, la suppression de cellules est préférable. Nous entendons par suppression, l'attribution d'un code de non-disponibilité sur l'enregistrement particulier touché par la contrainte de violation de confidentialité;

iv) si des cas d'unicité sont rencontrés au niveau univarié, regrouper des catégories. Examiner quelques distributions multivariées de variables jugées "sensibles" et effectuer les suppressions jugées nécessaires.

Un problème auquel nous avons dû faire face est relié à la méthode d'identification par appariement. En effet, un autre fichier de microdonnées, sous-ensemble de la population du recensement, a été produit. Ce fichier provenait de l'enquête sur la santé et les limitations d'activités (ESLA). De manière à restreindre tout potentiel d'identification par appariement entre les deux fichiers de microdonnées, la ventilation des variables communes aux deux fichiers a été de beaucoup réduite dans le fichier de l'ESLA. De plus les régions

géographiques identifiées dans le fichier de l'ESLA ont été restreintes.

4.3 Cas particulier: le revenu

Le fichier de microdonnées des ménages ne contient pas que des variables de nature qualitative; en effet, des variables telles le revenu et ses différentes composantes et le montant du loyer ou de l'hypothèque par exemple, proposent une valeur numérique exacte.

Nous nous attarderons ici au problème que nous a causé la variable revenu et ses différentes composantes (revenu d'investissement, de retraite, d'emploi autonome, d'emploi non autonome et autres revenus). Nous pouvons facilement imaginer le potentiel d'identification si un millionnaire ou une personne ayant vécu une faillite personnelle est sélectionné dans l'échantillon. Le croisement de seulement quelques caractéristiques telles l'origine ethnique, la profession et la géographie pourrait permettre d'identifier un individu ou ménage particulier.

Pour résoudre ce problème, nous avons considéré 3 approches:

i) imputation déterministe:

des limites supérieure et inférieure de revenu sont fixées et les valeurs du revenu et de ses composantes qui excèdent ces limites sont modifiées par le montant de ces limites. Les composantes du revenu sont ensuite ajustées de façon proportionnelle de manière à ce que leur somme totalise le revenu total ajusté. Le biais introduit est toutefois assez important.

ii) permutation aléatoire:

cette approche consiste à permuter aléatoirement les composantes du revenu pour tous les enregistrements pré-définis "à hauts risques". Étant donnée la très faible quantité d'enregistrements problèmes, le brouillage ne peut être effectué à l'intérieur de chaque région géographique identifiée dans le fichier. Ainsi, des groupes de brouillage sont déterminés par région de pondération, c'est-à-dire qu'un regroupement des régions géographiques avec poids semblables est effectué. Le revenu total est ensuite déduit en sommant les composantes du revenu modifiées. Cette méthode diminue de beaucoup le biais au niveau canadien par rapport à l'imputation déterministe. Toutefois, au niveau des RMR, les biais observés peuvent être assez importants étant donné le peu de cas à mélanger.

iii) addition d'un bruit aléatoire:

cette approche, décrite dans Kim (1986), consiste à ajouter un bruit aléatoire à chaque composante du revenu des individus définis à "hauts risques", puis à appliquer une transformation linéaire au revenu brouillé. Cette méthode, en plus d'être non biaisée lorsqu'on considère les données non pondérées, conserve la structure de corrélation des variables.

Lorsque nous ajoutons un bruit aléatoire, nous observons toujours de hauts et bas revenus dans le fichier. Nous nous demandons alors quelle quantité de variabilité est nécessaire pour assurer la protection face à l'identification. Lorsque nous mélangeons aléatoirement les revenus à hauts risques, un individu donné ne reçoit généralement pas son revenu. Si un millionnaire a été sélectionné, son revenu est imputé à quelqu'un d'autre. Ainsi, un utilisateur ne peut pas, à partir de l'information combinée, identifier un individu particulier dans la population.

Toutefois, non seulement le Bureau doit-il diminuer tout risque d'identification, mais il doit être perçu par le public comme ayant satisfait cette condition de protection. Or, même si la troisième méthode offrait un biais pondéré moindre, la première méthode, soit l'imputation déterministe, rencontrait mieux la condition et a donc été utilisée. De plus, cette méthode permettait de diminuer considérablement certains problèmes opérationnels.

Dans l'avenir, nous prévoyons poursuivre la recherche afin d'arriver à une méthode qui nous permettra d'allier une bonne protection des unités à une qualité accrue des données résultantes.

5. CONCLUSION

Il est donc assez difficile de contrer les risques d'identification d'un fichier de microdonnées. Tellement de facteurs sont impliqués tels la taille de l'échantillon utilisée, les unicités, l'identification géographique, la quantité de variables diffusées.

Dans le cas des unicités, le cas particulier proposé en guise d'application était relativement facile à résoudre étant donnée la connaissance de la population: recensement versus échantillon du recensement. Toutefois en pratique, la plupart des fichiers de microdonnées proviennent d'enquêtes par échantillonnage pour lesquelles nous ne possèdons aucun compte de population. Dans ce cas, certains auteurs ont proposé des méthodes d'estimation des unicités par modélisation (Bethlehem, Keller, Pannekoek 1988).

Étant donnés les futurs et présents développements technologiques, de plus en plus d'efforts seront portés dans l'avenir, vers le développement de méthodes pour contrer l'identification et la divulgation dans le cas d'échantillon de microdonnées. Une équipe de recherche vient tout juste d'être créée à Statistique Canada; elle examinera entre autres le problème de la confidentialité dans le cas de microdonnées et de tableaux de macrodonnées.

RÉFÉRENCES

1. Bethleem,J.G., Keller,W.J., Pannekoek,J. (1988), "Disclosure Control of Microdata", Netherlands Central Bureau of Statistics, 27 pages.

2. Denis, J. (1988), "Bandes de microdonnées du recensement -Fichier des ménages, problème du revenu", Division des méthodes d'enquêtes sociales, Statistique Canada.

3. Kim, J. (1986), "A Method for Limiting Disclosure in Microdata Based on Random Noise and Transformation", ASA Proceedings of the Survey Research, Methods Section, 370-374.

4. "Report on Statistical Disclosure and Disclosure Avoidance Techniques" (1978), Statistical Policy Working Paper 2, U.S. Department of Commerce, Office of Federal Statistical Policy and Standards, 23-31.

SESSIONS SUR LES SÉRIES CHRONOLOGIQUES

Modélisation multivariée du courant de Gaspé

G. BÉGIN, A. LATOUR, Université du Québec à Montréal
J.-P. CHANUT, M. I. EL-SABH, Université du Québec à Rimouski.

RÉSUMÉ

Plusieurs applications des séries chronologiques se retrouvent en économie où les effets saisonniers sont observés à des fréquences qui sont des multiples de l'intervalle d'observation. Par exemple, un taux mensuel de chômage peut présenter une période de 12 mois. Ce n'est cependant pas le cas en sciences naturelles. Il est bien connu que les courants marins sont influencés par les marées et que ces dernières sont déterminées par la position des astres. Les périodes impliquées ne sont pas nécessairement des multiples entiers de l'intervalle de temps entre deux observations, ce qui complique la modélisation. Trois stations ont été utilisées afin de recueillir des données relatives au courant de Gaspé. Un modèle pour la vitesse du courant est proposé. On y retrouve une composante déterministe faisant intervenir onze fréquences naturelles et une composante stochastique représentée par un processus ARIMA multivarié.

Mots clés: *Séries chronologiques multivariées, régression harmonique , modélisation, analyse des courants marins*

1. INTRODUCTION

Munis d'appareils permettant l'enregistrement automatique et continu (ou quasi-continu), les océanographes désirent améliorer leur connaissance de l'évolution temporelle et spatiale de diverses caractéristiques biologiques, chimiques et physiques des milieux marins. Ces chercheurs recueillent une très grande quantité d'informations nécessitant ainsi l'utilisation de nouvelles techniques d'analyse de données. Dans la présente étude, nous utiliserons les techniques relatives à l'analyse des séries chronologiques.

ANDERSON (1971) et FULLER (1976) donnent les résultats de base de l'analyse des séries chronologiques univariées en considérant l'approche temporelle et l'approche fréquentielle. BLOOMFIELD (1976) et BRILLINGER (1981) explicitent les résultats de l'approche fréquentielle. Par contre, la véritable utilisation des modèles univariés est due à BOX et JENKINS (1976). Les logiciels SAS et BMDP permettent de faire une modélisation dans le cas univarié.

Pour le cas multivarié, QUENOUILLE (1968) et PRIESTLEY (1981) développent les concepts théoriques de base. C'est grâce aux résultats de TIAO et BOX (1981) et de TIAO et TSAY (1983) que l'application multivariée est devenue possible.

En océanographie, l'approche habituelle fait appel à l'analyse spectrale pour tenter d'isoler les périodes agissant sur une série d'observations (WHITTLE, 1954; PLATT et DENMAN, 1975; ROBERTS, 1982; RAO et YU, 1986; et plusieurs autres). Par contre, peu d'auteurs considèrent l'approche temporelle. Dans le cas univarié, CHANUT et al. (1987, 1988) utilisent cette approche pour enrichir les résultats de l'analyse spectrale. Finalement, nous n'avons trouvé aucun article combinant les deux approches pour le cas des séries chronologiques multivariés.

Les séries de données océanographiques sont généralement de type pseudo-périodique c'est-à-dire qu'elles se caractérisent par une composante déterministe et une composante stochastique. La partie déterministe présente des fréquences d'origines diverses. Les fréquences induites par la lune et le soleil, constituent un groupe important de fréquences agissant sur les observations de caractères physiques. Également des fréquences dues aux oscillations et aux réflections des niveaux d'eau dans les baies engendrées par les marées sont présentes. La profondeur, la longueur, la largeur et le relief du fond de l'estuaire produisent aussi des fréquences influençant l'environnement marin. De plus, on retrouve des fréquences produites par des divers phénomènes tels que la remontée d'eaux profondes, les gets côtiers, les tourbillons, la décharge de rivières estuariennes, etc... Finalement, on peut mentionner que les vents peuvent produire des fréquences mais que ces dernières ont une période inférieure à cinq minutes (MCLELLAN, 1975).

L'objectif de notre étude est de suggérer un modèle pour le courant de Gaspé. Ce courant, prenant origine à la hauteur du Saguenay, longe la côte de la péninsule gaspésienne et se jette dans le golfe du Saint-Laurent. L'étude du courant de Gaspé suscite un grand intérêt en océanographie puisqu'elle permet d'améliorer la connaissance des caractéristiques biologiques du golfe. En effet, ce courant joue un rôle important dans le contrôle de la production biologique et dans la distribution de l'ichtyoplancton (oeufs et larves de divers organismes marins) (voir STEVEN (1974); SINCLAIR et al. (1976); SÉVIGNY et al. (1979); EL-SABH et al. (1984); DE LAFONTAINE et al. (1984)).

C'est durant l'été 1978 que l'enregistrement des données a eu lieu grâce à la participation conjointe de l'Institut Océanographique de Bedford (Darmouth, N.-É., Canada) et l'Université du

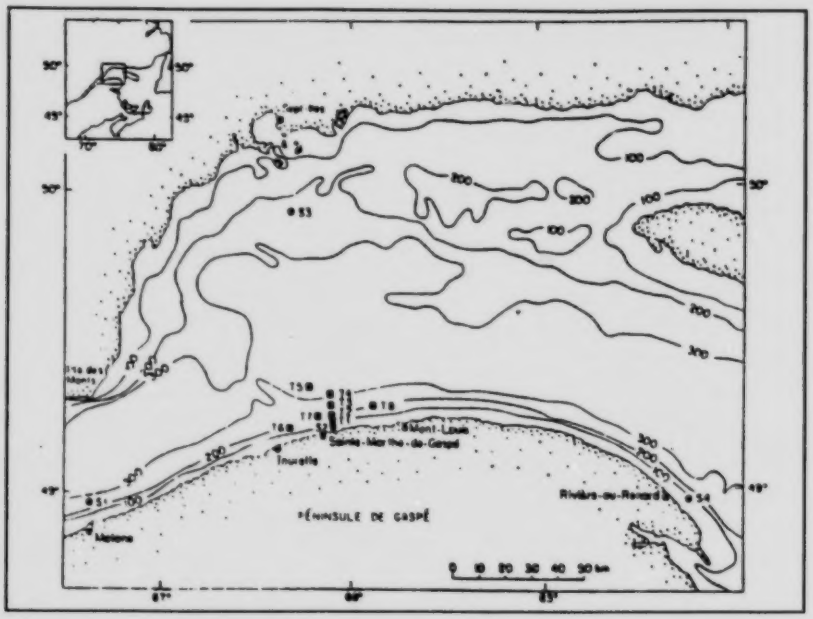


Figure 1: *Localisation des stations d'échantillonnage (Source: Benoît, 1980)*

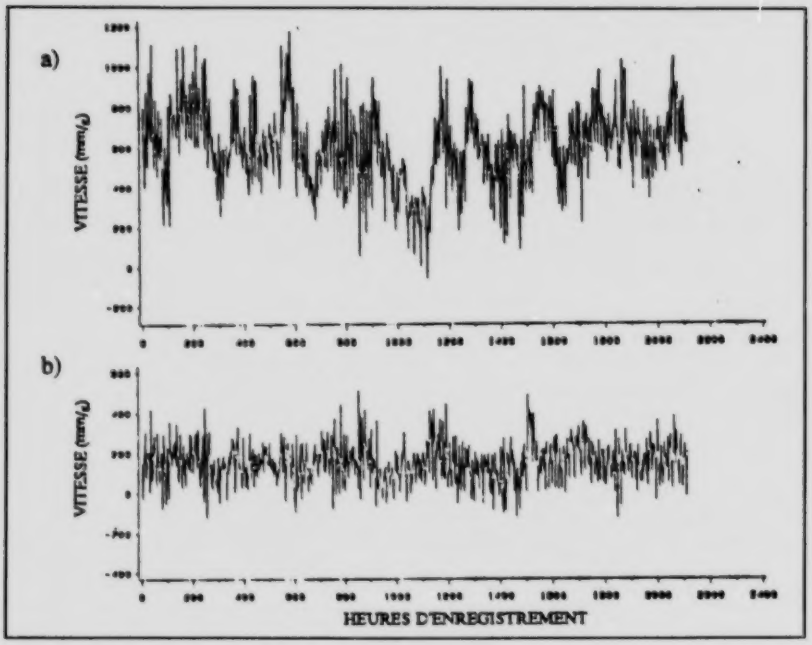


Figure 2: Vitesse observée à la station T_2: a) Composante Est et b) Composante Nord

Québec à Rimouski (Québec, Canada). Le but général de l'étude était, à cette époque, l'amélioration de la compréhension de la dynamique de la circulation du courant de Gaspé.

L'objectif plus spécifique de notre présent travail est de modéliser la vitesse du courant de Gaspé en utilisant la théorie des séries chronologiques multidimensionnelles. Pour ce faire, nous considérerons la composante azimut Est de la vitesse à trois stations différentes et à une profondeur commune de vingt mètres. Suite à l'analyse spectrale qui nous permettra d'isoler les principales fréquences astronomiques, nous appliquerons la régression

harmonique sur nos observations. Dans un deuxième temps, nous appliquerons la théorie des séries chronologiques multivariées sur les résidus obtenus de la régression harmonique.

2. Description générale

Les données recueillies relativement au courant de Gaspé proviennent de trois stations d'échantillonnage, "T_2", "T_4" et "T_5", voir figure 1. Elles ont été prises du 22 juin au 15 septembre 1978. La vitesse et la direction du courant, de même que la pression, la température et la conductibilité de l'eau ont été enregistrées à toutes les vingt minutes durant cette période. Nous ne retenons que la composante Est de la vitesse, puisqu'il s'agit de la composante la plus importante (voir figure 2). Ayant une quantité énorme de données, (plus de 6 000 observations par station), nous considérons une moyenne horaire afin de pouvoir en faire le traitement informatique.

Station	Nombre d'observations	Moyenne	Écart-type	Min.	Max.
T_2	2115	594	186	-64	1173
T_4	2322	440	173	-75	974
T_5	2335	203	157	-342	735

Tableau 1: *Statistiques descriptives pour les trois stations*

	T_2	T_4	T_5
T_2	1.000	0.086	0.266
T_4	0.086	1.000	0.041
T_5	0.266	0.041	1.000

Tableau 2: *Corrélations croisées de délai 0.*

Composante harmonique	*Symb.*	*Fréquence (cycles/hre)*	*Période*
Semi-diurnes			
lunaire principale	M_2	0.08051140	12.42
solaire principale	S_2	0.08333333	12.00
lunaire elliptique majeure	N_2	0.07899925	12.66
luni-solaire déclinationnelle	K_2	0.08356149	11.97
lunaire elliptique mineure	L_2	0.08202355	12.19
	η_2	0.08507364	11.75
Diurnes			
luni-solaire	K_1	0.04178075	23.93
lunaire principale	O_1	0.03873065	25.82
solaire principale	P_1	0.04155259	24.07
lunaire déclinationnelle mineure	Oo_1	0.04483084	22.31
	$2Q_1$	0.03570635	28.01
	α_1	0.03439657	29.07
	υ_1	0.04634299	21.59
De longue période			
lunaire semi-mensuelle	M_f	0.00305008	327.86
lunaire mensuelle	M_m	0.00151215	661.30
	M_{sf}	0.00282193	354.37

Tableau 3: *Composantes harmoniques principales*

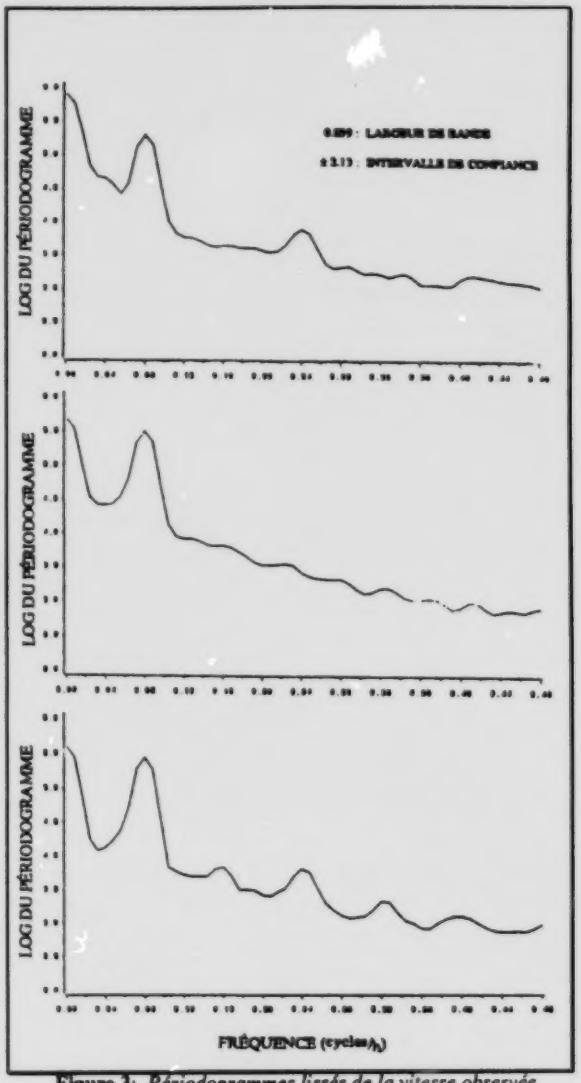


Figure 3: *Périodogrammes lissés de la vitesse observée à chacune des trois stations a) T_2, b) T_4 et c) T_5.*

3. REMARQUES PRÉLIMINAIRES

Au tableau 1 nous présentons quelques statistiques descriptives reliées à la vitesse du courant observée aux trois stations. La moyenne décroît de T_2 à T_5. Ceci est normal puisque la station T_2 est au milieu de l'aire où le courant est très fort, et que la force du courant décroît de T_2 à T_4 à T_5.

Pour la modélisation, nous considérons les 2,115 observations enregistrées simultanément aux trois stations. Au tableau 2 nous donnons les corrélations entre les vitesses observées.

Il est intéressant de noter que la plus grande corrélation est entre les stations T_2 et T_5, et que ces deux stations sont les plus éloignées l'une de l'autre.

Des contraintes d'ordre pratique nous incitent à considérer des fréquences qui ne sont pas inférieures à

$$f_0 = \frac{2}{N\Delta_t},$$

ou supérieures à

$$f_q = \frac{1}{2\Delta_t},$$

où N est le nombre d'observations et Δ_t est l'intervalle entre deux observations. Ces deux bornes sont respectivement 0.00009 et 0.5 $^{\text{cycles}}/_{\text{hre}}$.

Le tableau 3 présente les principales composantes harmoniques jugées intéressantes a priori. Elles sont de trois types: six semi-diurnes, sept diurnes et trois de longues périodes. Nous voyons que K_1 et P_1, tout comme S_2 et K_2, sont des fréquences très près l'une de l'autre. Le critère de Rayleigh nous indique que deux fréquences f_1 et f_2 ne peuvent pas être distinguées si $|f_1 - f_2| \times N \leq 1$ cycle. Cela signifie qu'il faudrait plus de 4 000 heures d'observation pour différencier ces fréquences.

4. LE MODÈLE

Relativement à la mise sur pied de notre modèle, la première étape consiste à choisir un ensemble de composantes harmoniques. Elles détermineront la composante déterministe du modèle. La deuxième étape est l'expression de l'erreur en termes d'un modèle ARIMA multivarié.

Soit

$$\underset{\sim}{X}_t = [X_{T_2}(t), X_{T_4}(t), X_{T_5}(t)]'$$

le vecteur de la vitesse observée à chacune des stations au temps t. Nous supposons que

$$\underset{\sim}{X}_t = \underset{\sim}{\varphi}_t + \underset{\sim}{Z}_t, \qquad (1)$$

où $\underset{\sim}{\varphi}_t$ est la composante déterministe et $\underset{\sim}{Z}_t$ est la composante stochastique. Ces composantes sont exposées en détails aux sections 4.1. et 4.2. respectivement.

4.1. LA COMPOSANTE DÉTERMINISTE

Il s'agit de déterminer les fréquences auxquelles semblent être reliées nos séries. L'analyse spectrale peut être utilisée afin d'explorer l'ensemble des fréquences intéressantes. La figure 3 nous montre les périodogrammes lissés de la vitesse observée à chacune des stations.

On détecte un maximum local dans le voisinage de 0.08. Cette valeur correspond à des composantes ayant une période d'environ douze heures. Dans le premier et dans le troisième périodogrammes, nous notons aussi un autre maximum local dans le voisinage de 0.24 , ce qui correspond à des composantes ayant une période d'environ quatre heures.

Nous avons utilisé le programme WTLVAN (WATER LEVEL ANALYSIS), écrit par M. FOREMAN (voir FOREMAN (1977)) afin de mieux cerner l'ensemble des fréquences "naturelles" pouvant se manifester dans la série. Ce programme prend en considération 120 fréquences astronomiques, la date et le lieu d'où proviennent les observations. Il peut donc être utilisé afin de produire une liste des fréquences les plus importantes pouvant se manifester dans de telles séries. La composante déterministe du modèle est

$$\varphi_{T_j}(t) = \mu_j + \sum_{i=1}^{k} A_{T_j,i} \sin(2\pi f_{T_j,i} t + \theta_{T_j,i}), \qquad (2)$$

$1 \leq t \leq N$, et $j \in J = \{2, 4, 5\}$ ou, d'une manière équivalente

$$\varphi_{T_j}(t) = \mu_j + \sum_{i=1}^{k} \{\alpha_{T_j,i} \sin(2\pi f_{T_j,i} t) + \beta_{T_j,i} \cos(2\pi f_{T_j,i} t)\} \qquad (3)$$

$1 \leq t \leq N$, et $j \in J$.

La fonction $\varphi_{T_j}(t)$ est définie en termes de $2k + 1$ paramètres. Les termes $\sin(2\pi f_{T_j,i} t)$ et $\cos(2\pi f_{T_j,i} t)$ peuvent être considérés comme des variables explicatives dans un modèle de régression. La statistique C_p de Mallows (DRAPER et SMITH, 1981) est utilisée afin de choisir le meilleur ensemble de variables explicatives. D'un point de vue pratique, nous utilisons la procédure 9R de BMDP, (Dixon, 1983). Puisque cette procédure ne peut considérer plus de vingt-sept variables indépendantes, 13 composantes harmoniques ont été prises en considération. Nous avons retenu les fréquences suivantes:

$\{M_m, M_{sf}, \alpha_1, 2Q_1, O_1, K_1, O_{o1}, N_2, M_2, S_2, \eta_2, \upsilon_1, L_2\}$.

où $\{\underset{\sim}{Z}_t\}$ est un processus stochastique. Pour cette composante aléatoire, un modèle est construit en suivant l'approche basée sur les travaux de TIAO et BOX (1981), de même que TIAO et TSAY (1983). Le logiciel SCA a été mis au point dans cette optique. Pour les détails, voir LIU, HUDAK et al. (1986). Dans la représentation (1), $\{\underset{\sim}{Z}_t\}$ peut être un processus non stationnaire (voir CHANUT et al. (1987)). Afin d'obtenir un processus stationnaire, un différence régulière d'ordre 1 peut être appliquée:

$$\underset{\sim}{\omega}_t = \nabla \underset{\sim}{Z}_t = \underset{\sim}{Z}_t - \underset{\sim}{Z}_{t-1} \qquad (4)$$

Le processus $\{\underset{\sim}{\omega}_t\}$ s'exprime alors sous la forme

$$\underset{\sim}{\omega}_t = \Phi_1 \underset{\sim}{\omega}_{t-1} + \ldots + \Phi_p \underset{\sim}{\omega}_{t-p} + \underset{\sim}{a}_t - \Theta_1 \underset{\sim}{a}_{t-1} \ldots - \Theta \underset{\sim}{a}_{t-q} \qquad (5)$$

où Φ_j, $1 \le j \le p$, et Θ_k, $1 \le k \le q$, sont des matrices 3x3, et $\underset{\sim}{a}_t$ est un bruit blanc de moyenne nulle et de matrice de covariance $\Sigma_{\underset{\sim}{a}}$. Il peut être préférable d'écrire ce modèle sous la forme

$$\Phi(B)\underset{\sim}{\omega}_t = \Theta(B)\underset{\sim}{a}_t \qquad (6)$$

où B est l'opérateur de décalage arrière ($B\underset{\sim}{X}_t = \underset{\sim}{X}_{t-1}$), tandis que $\Phi(B)$ et $\Theta(B)$ sont des matrices 3x3 dont les éléments sont des polynômes en B.

5. RÉSULTATS

Les valeurs estimées des paramètres de la composante déterministe sont présentées au tableau 4. Les coefficients de détermination, R^2, sont respectivement 0.27, 0.39 et 0.44 pour les stations T_2, T_4 et T_5. Ce qui signifie que les onze fréquences introduites permettent d'expliquer entre 27% et 44% de la variation expliquée. Les graphiques de la vitesse résiduelle à chacune des stations sont donnés à la figure 4.

On constate que ces séries ne sont pas stationnaires. Une différence d'ordre 1 est alors appliquée à ces vitesses résiduelles. Une telle différenciation avait aussi été nécessaire dans une étude antérieure (CHANUT et al., 1987). Les autocorrélations échantillonnales croisées et étendues ("*Extended Sample Cross-Correlations*") ont été calculées avec le logiciel SCA, et le modèle suivant a été retenu:

$$(I - \Phi_1 B - \Phi_2 B^2 - \Phi_3 B^3 - \Phi_6 B^6 - \Phi_{12} B^{12} - \Phi_{13} B^{13})\underset{\sim}{\omega}_t$$
$$= (I - \Theta_1 B - \Theta_2 B^2 - \Theta_3 B^3)\underset{\sim}{a}_t \qquad (5)$$

Les coefficients estimés de ce modèle sont donnés au tableau 5. Un signe "+" ou un signe "–" indique que la valeur absolue correspondante est supérieure à deux écarts-types; un signe "+" pour les valeurs positives et un signe "–" pour les valeurs négatives. L'ajustement de ce modèle est jugé très satisfaisant.

	Amplitude			*Phase*			*Rapport des amplitudes* M_2 : 100		
	T_2	T_4	T_5	T_2	T_4	T_5	T_2	T_4	T_5
M_m	24.33	16.42	25.91	- 0.69	0.43	- 1.17	23.19	12.25	20.59
M_{sf}	61.78	36.13	48.86	0.04	- 0.30	- 1.12	58.89	26.96	38.83
α_1	12.02	6.90	5.69	- 1.04	- 0.21	0.27	11.46	5.15	4.52
$2Q_1$	8.63	4.27	2.21	- 1.16	0.29	- 1.28	8.23	3.19	1.76
O_1	27.49	14.66	10.26	0.42	0.82	0.40	26.21	10.94	8.15
K_1	30.54	21.00	14.43	0.49	- 1.31	1.23	29.11	15.67	11.47
N_2	19.34	39.33	33.92	1.02	- 0.11	1.20	18.44	29.35	26.96
M_2	104.9	134.0	125.8	- 0.35	- 1.53	- 0.15	100.0	100.0	100.0
L_2	9.67	7.10	10.77	1.44	0.97	- 1.09	9.22	5.30	8.56
S_2	25.77	34.42	30.80	1.10	- 0.05	1.55	24.57	25.69	24.48
η_2	9.60	7.01	7.74	0.55	- 0.57	- 1.32	9.15	5.23	6.15

Tableau 4: *Paramètres estimés de la composante déterministe.*

Ordre	Φ				Θ				*Ordre*	Φ			
1	-.129	-.652	.842	. - +	.032	-.652	.852	. - +	6	-.172	.030	.005	- . .
	(.139)	(.395)	(.230)		(.141)	(.396)	(.232)			(.024)	(.026)	(.026)	
	-.176	1.181	-.188	. + .	-.160	1.218	-.230	. + .		.032	-.050	-.019	. - .
	(.127)	(.087)	(.147)		(.129)	(.085)	(.147)			(.021)	(.019)	(.020)	
	-.115	.402	.908	. + +	-.111	.354	1.109	. + +		-.055	.002	-.067	- . -
	(.108)	(.109)	(.052)		(.109)	(.108)	(.049)			(.019)	(.018)	(.019)	
2	.167	.383	-.785	. - -	.377	.340	-.921	+ - -	12	.089	.057	.098	+ + +
	(.125)	(.342)	(.198)		(.130)	(.351)	(.237)			(.022)	(.024)	(.026)	
	.086	-.416	.395	. - +	.132	-.219	.507	. - +		.033	.074	-.041	. + .
	(.124)	(.098)	(.134)		(.141)	(.101)	(.161)			(.021)	(.023)	(.026)	
	.021	-.469	-.334	. - -	-.002	-.457	-.227	. - -		.029	-.003	.210	. . +
	(.099)	(.094)	(.072)		(.112)	(.098)	(.075)			(.019)	(.020)	(.023)	
3	-.035	.237	.187	. . .	.108	.301	.198	. + .	13	.082	.089	-.222	+ . -
	(.096)	(.120)	(.115)		(.123)	(.137)	(.156)			(.035)	(.050)	(.059)	
	.027	.018	-.237	. . -	.019	-.055	-.310	. . -		.002	-.015	-.002	. . .
	(.067)	(.081)	(.070)		(.086)	(.102)	(.090)			(.023)	(.025)	(.044)	
	.054	.232	.127	. + +	.078	.121	.092	. . .		-.012	-.020	-.012	. . .
	(.060)	(.068)	(.061)		(.076)	(.083)	(.081)			(.023)	(.026)	(.029)	

Tableau 5: *Paramètres estimés du modèle ARIMA multivarié identifié pour la suite des vitesses résiduelles.*

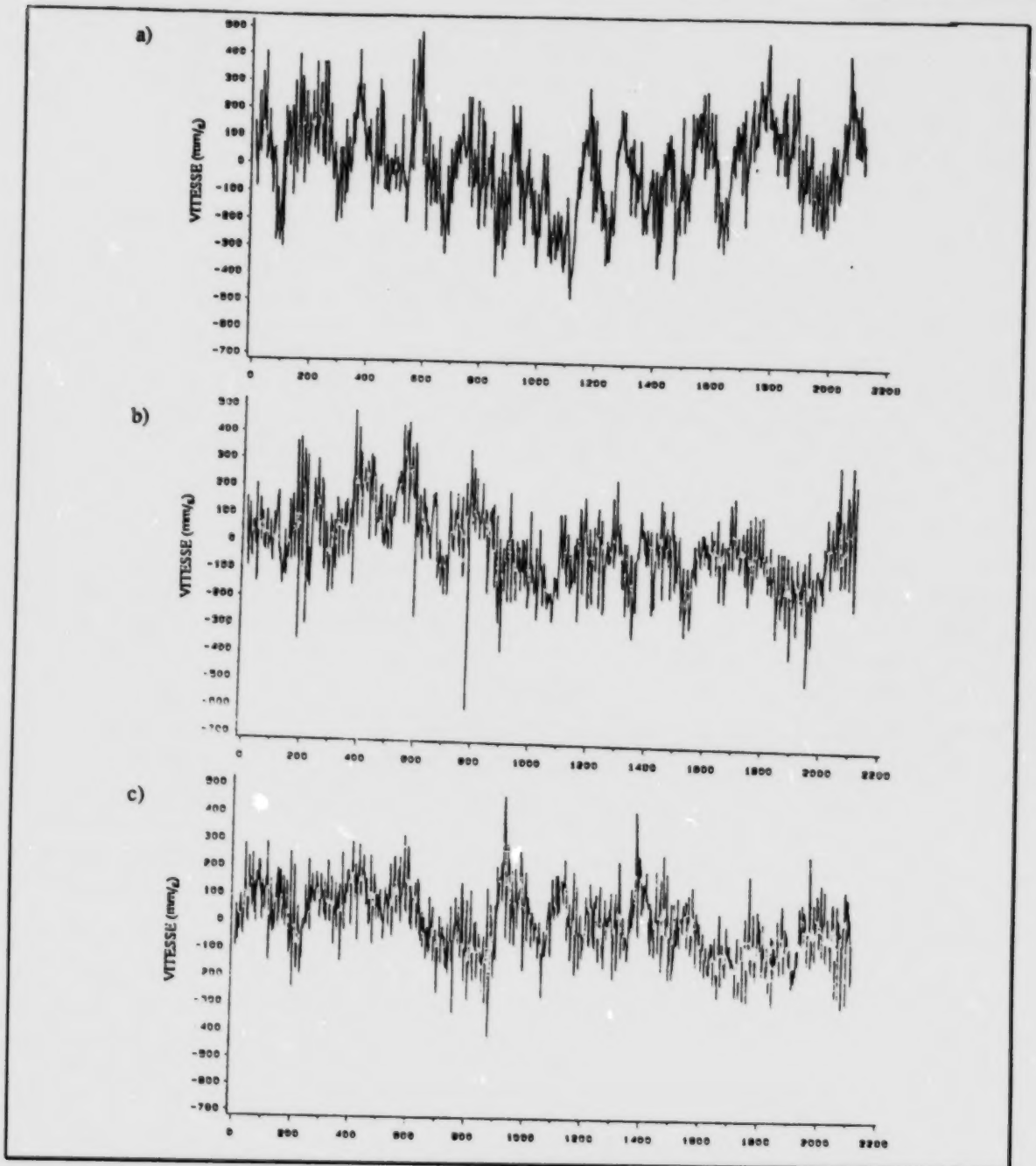


Figure 4: *Vitesse résiduelle pour chacune des stations: a)T_2, b)T_4 et c)T_5*

Appliquant la statistique C_p de Mallows, O_{o1} et υ_1 sont retirées de l'ensemble des "meilleures" variables explicatives. Ainsi, la composante déterministe fait intervenir les fréquences suivantes:

$$\{M_m, M_{sf}, \alpha_1, 2Q_1, O_1, K_1, N_2, M_2, S_2, \eta_2, L_2\}.$$

4.2. LA COMPOSANTE STOCHASTIQUE

Dans le modèle (1) nous supposons que

$$\underset{\sim}{X}_t = \underset{\sim}{\varphi}_t + \underset{\sim}{Z}_t .$$

6. Discussion

Tout comme CHANUT et al. (1987), l'utilisation de l'analyse spectrale nous a conduit à introduire des composantes harmoniques dont les fréquences sont dans le voisinage de 0.08 et de 0.24 $^{cycles}/_{hre}$. De telles composantes peuvent expliquer en partie le développement pseudo-périodique de nos séries. Cependant, nous avons mis sur pied un modèle ayant plus de composantes harmoniques que CHANUT et al.. Notre approche multivariée permet le traitement implicite d'une composante spatiale. De plus, l'ordre des opérateurs, autorégressif et de moyenne mobile, sont moindres que ceux introduits par CHANUT et al. dans leur modèle. Le modèle ajusté indique que la vitesse résiduelle différenciée du courant $\underset{\sim}{a}_t$ peut être expliquée en termes des valeurs antérieures de délai 1, 2, 3, 6 et 12, et en termes des termes d'erreur d'un délai variant entre 1 et 3.

Seulement deux éléments sur neuf sont différents de zéro dans la matrice Φ_{13}, mais nous ne pouvons retirer cette matrice du modèle. Certaines fréquences astronomiques ayant une valeur dans le voisinage de douze ou treize heures n'ont pas été introduites dans le modèle. Ceci peut expliquer les facteurs autorégressifs d'ordre 12 et 13.

Références

ANDERSON, T.W. (1971) *The Statistical Analysis of Time Series*, New York: Wiley and Sons, 704 pages.

BENOIT, J. (1980) *Variations temporelles et spatiales du courant de Gaspé*, Mémoire de maîtrise: Université du Québec à Rimouski, 108 pages.

BLOOMFIELD, P. (1976) *Fourier Analysis of Time Series: An Introduction*, New York: Wiley and Sons, 538 pages.

BOX, G.E.P. et JENKINS, G.M. (1976) *Time series analysis: forecasting and control* (édition revisée), San Francisco: Holden-Day, 575 pages.

BOX, G.E.P. et TIAO, G.C. (1977) A canonical analysis of multiple time series, *Biometrika*, vol. 64, no 2, pp. 355-365.

BRILLINGER, D.R. (1981) *Time series: Data Analysis and Theory*, San Francisco: Holden-Day, 540 pages.

CHANUT, J.P. et al. (1987) A stochastic modelling of tidal current measurements, *Stochastic Hydrology*, 4, pp. 73-86.

CHANUT, J.P. et al. (1988) Modelling the natural and anthropogenic variations of the St.Lawrence water level, *Natural and Man-Made Hazards*, pp. 377-394.

DE LAFONTAINE, Y. et al. (1984) Structure océanographique et distribution spatio- temporelle d'oeufs et de larves de poissons dans l'estuaire maritime et la partie ouest du Golfe St-Laurent, *Sciences et Techniques de l'Eau*, vol. 17, pp. 43-50.

DIXON, W.J. et al. (1983) *BMDP Statistical Software* (édition revisée), University of California.

DRAPER, N. R. et SMITH, H. (1981) *Applied Regression Analysis* (2ième édition), New York: Wiley & Sons, 709 pages.

FOREMAN, M.G.G. (1977) *Manual for Tidal Heights Analysis and Prediction*, Pacific Marine Science Report 77-10; Institute of Ocean Sciences, Patricia Bay, Victoria (B.C.), 101 pages.

FULLER, W.A. (1976) *Introduction to statistical Time Series*, New York: Wiley and Sons, 470 pages.

LIU, L.M., HUDAK, G.B. et al. (1986) *Scientific Computing Associates*, Illinois.

McLELLAN, H.J. (1975) *Elements of physical Oceanography*, Pergamon Press Inc., 150 pages.

PLATT, T. et DENMAN, K.L. (1975) Spectral Analysis in Ecology, *Annal Review of ecology and systematics*, vol. 6, pp. 189-210.

PRIESTLEY, M. B. (1981) *Spectral Analysis and Time Series*; Volume 1: *Univariate series*; Volume 2: *Multivariate series, prediction and control*, New York: Academic Press, 890 pages.

QUENOUILLE, M.H. (1968) *The analysis of multiple Time-Series* (2ième édition), New York: Hafner publishing company, 105 pages.

RAO, A.R. et YU, G.H. (1986) Detection of nonstationarity in hydrologic Time Series, *Management Science*, vol. 32, no 9, pp. 1206-1217.

ROBERTS, P.J.W. (1982) Analysis of current meter data for predicting pollutant dispersion, *Time Series Methods in Hydrosciences*, pp. 351-361.

SÉVIGNY, J.-M. et al. (1979) Summer Plankton distribution associedwith the physical and nutrient properties of the northwestern Gulf of St.Lawrence, *Journal of the Fisheries Research Board of Canada*, 36, pp. 187-203.

SINCLAIR, M., EL-SABH, M. I. et BRINDLE, J.-R. (1976) Seaward nutrient transport in the lower St.Lawrence Estuary, *Journal of the Fisheries Research Board of Canada*, 33, pp. 1271-1277.

STEVEN, D. M. (1974) *Primary and secondary production in the Gulf of St.Lawrence*, McGill University: Marine Science Center, rapport no.26.

TIAO, G.C. et BOX, G.E.P. (1981) Modelling Multiple Time Series With Applications, *Journal of the American Statistical Association: Theory and Methods Section*, vol. 76, no 376, pp. 802-816.

TIAO, G.C. et TSAY, R.S. (1983) Multiple Time Series Modeling and Extented Sample Cross-Correlations, *Journal of Business & Economic Statistics*, vol. 1, no 1, pp. 43-56.

WHITTLE, P. (1954) The statistical analysis of a seiche record, *Journal of Marine Research*, vol. 13, pp. 76-100.

Étude comparative de modèles chronologiques décrivant la réponse cardiaque en situation réelle de travail *

Carole Brabant (1), Alain Latour (2), Jocelyne Ferraris (1), Donna Mergler (1)

Résumé

Dans le but d'extraire les informations contenues dans les enregistrements de la fréquence cardiaque en situation de travail, nous poursuivons notre démarche modélisante suivant l'approche de Box et Jenkins. Disposant de plusieurs séries décrites en termes 1- de structure de corrélation temporelle (modèles ARIMA), 2- de comportement relié aux caractéristiques de l'activité (fonctions de transfert et modèles d'intervention), nous proposons l'analyse comparative des modèles ajustés pour explorer la variabilité hebdomadaire et saisonnière de la réponse cardiaque. Les perspectives d'interprétation de cette démarche sont illustrées à partir de quelques profils de fréquence cardiaque.

Mots clés: séries chronologiques, fonctions de transfert, modèles d'intervention, rythme cardiaque, variabilité.

(1) Groupe de Recherche-Action en Biologie du Travail (associé à l'Institut de Recherche en Santé et en Sécurité du Travail). Département des sciences biologiques, Université du Québec à Montréal, case postale 8888, succursale A, Montréal (Québec) H3C 3P8.

(2) Département de mathématiques et informatique, Université du Québec à Montréal, case postale 8888, succursale A, Montréal (Québec) H3C 3P8.

* Travaux subventionnés par le Conseil de Recherche en Sciences Humaines (CRSH) et par l'Institut de Recherche en Santé et en Sécurité du Travail (IRSST).

1. Introduction

Une récente étude menée auprès de travailleuses de buanderie rapportant des symptômes d'inconfort thermique malgré le respect des normes d'exposition à la chaleur, a permis de mettre en évidence une surcharge cardiaque importante, surpassant les limites recommandées dans la littérature[1]. Ce constat va dans le sens d'une sous-estimation de la charge de travail qui implique une série de contraintes gestuelles et posturales particulièrement exigeantes pour le système cardiovasculaire. Il soulève également la pertinence du développement de critères normatifs propres aux travaux sédentaires, reflétant mieux la charge physiologique résultant de l'astreinte cardiaque, possiblement corollaire de l'inconfort thermique et indicatrice d'éventuels désordres chroniques.

La fréquence cardiaque est un témoin global du niveau d'activité de travail, reflétant non seulement le coût énergétique de la contraction musculaire, mais également les réactions de l'organisme face aux variations de l'ambiance physique (température, niveaux de bruit...), aux émotions et aux agressions diverses[2]. La très grande sensibilité de la réactivité cardiaque est un des attributs majeurs de ce paramètre en tant que variable intégratrice des contraintes subies par un organisme. Par ailleurs, cette sensibilité limite les possibilités d'identification et de quantification des sources de variabilité intrinsèque et extrinsèque qui se superposent à la variabilité inter-individuelle. Dans ce contexte, il devient souvent nécessaire d'élaborer des protocoles couvrant un maximum de variables explicatives du comportement cardiaque, incluant la temporalité. Or, la multiplication des données obtenues lors d'enregistrements répétés et idéalement continus de la fréquence cardiaque complexifie la manipulation des résultats et nécessite des outils statistiques adéquats.

Les techniques de modélisation chronologique, utilisées en complément de l'analyse de la fréquence cardiaque à travers différents indices d'astreinte et descripteurs mathématiques, permettent de ramener plusieurs profils en une série d'équations paramétriques simples (famille ARIMA) et de mettre en relation ces profils avec d'autres variables (fonctions de transfert et modèles d'intervention), tout en conservant le caractère temporel. Cette démarche, appliquée à plusieurs séries couvrant en partie la variabilité intra-individuelle (6 jours par travailleuse) et inter-contextuelle (2 saisons par travailleuse, plusieurs tâches et cycles de travail) nous conduit à la comparaison des différents modèles pour explorer les sources et l'étendue de la variabilité. L'étude de la variabilité et de la stabilité relative de la réactivité cardiaque en situation réelle de travail s'insère dans un objectif beaucoup plus global de validation du critère cardiaque et d'identification des composantes du travail les plus contraignantes en vue d'orienter les interventions ergonomiques prioritaires.

2. Nature des données

L'étude a porté sur 11 travailleuses d'une buanderie industrielle. Chaque travailleuse a été suivie à raison de 3 jours complets d'une même semaine. Le protocole a été répété l'été, lorsque la température des lieux de travail est chaude et l'hiver, lorsque l'ambiance thermique est plutôt confortable. On dispose au total de 66 jours-travailleuses répartis sur les 2 saisons.

On a récolté des données en regard de 5 volets: 1- l'environnement thermique; 2- les caractéristiques individuelles; 3- l'activité de travail; 4- la symptomatologie et 5- la réponse physiologique. De l'ensemble des variables couvrant les différents volets de l'étude, on ne retiendra que la fréquence cardiaque et le nombre de mouvement(s) du bras dominant enregistrés toutes les 30 secondes pendant toute la journée de travail (Vitalog PMS-8), de même que le poste occupé, extrait de l'observation continue de l'activité de travail (Sharp PC 1500A).

3. Modélisation statistique

La démarche modélisante repose sur 3 objectifs complémentaires soit: 1- décrire l'évolution temporelle de la fréquence cardiaque; 2- quantifier l'impact des bras sur l'accélération cardiaque; et 3- hiérarchiser l'impact des différentes tâches sur le comportement cardiaque. Ces objectifs spécifiques convergent tous vers une meilleure connaissance de la réactivité cardiaque en situation réelle de travail, ce qui passe nécessairement par l'identification des sources et de l'étendue de la variabilité aussi bien inter qu'intra-individuelle. L'analyse de la variabilité et de la stabilité relative des réponses cardiaques repose sur la comparaison des résultats de la modélisation[3], incluant les procédures d'analyse univariée appliquées aux séries chronologiques[4] et deux approches dérivées faisant intervenir des variables extrinsèques, les fonctions de transfert et les modèles d'intervention[5]. Quelle que soit l'approche, la modélisation s'effectue en 3 étapes distinctes, l'identification du modèle, l'estimation des paramètres, et la validation.

L'identification du modèle s'inspire de l'examen des propriétés statistiques de la série, notamment l'allure et la valeur estimée des autocorrélations totales, partielles et des corrélations croisées dans le cas des procédures bivariées. La validité de ces diverses estimations, dérivées de la notion de corrélation, est conditionnée par la stationnarité des séries, concept qui qualifie les séries de moyenne et de covariance constantes, de sorte que $E[X_t] = \mu$ et que $Cov[X_t, X_{t+k}] = \gamma_k$ quelque soit t. La stationnarité peut être assurée en appliquant une ou plusieurs différenciation(s) à la série originale, sans limite interprétative.

L'estimation des paramètres s'effectue via des algorithmes itératifs complexes qui permettent d'optimiser l'ajustement des paramètres suivant le principe du maximum de vraisemblance.

Le premier pas de l'étape de validation est la signification des paramètres, vérifiée par un test-t. Le second pas est l'examen des résidus qui se doivent d'être des variations aléatoires et dépourvues de toute structure de corrélation inexpliquée par le modèle ajusté, pour correspondre à la définition d'un bruit blanc. Finalement la comparaison des valeurs prédites par le modèle et des valeurs observées permet de juger l'adéquation du modèle (l'erreur relative), donc indirectement sa validité.

3.1 Les modèles chronologiques

Les modèles que nous privilégions pour décrire la structure de corrélation de la fréquence cardiaque, en regard du premier objectif, font partie de la famille désignée par l'acronyme ARIMA (*Autoregressive Integrated Moving Average*). Ces modèles, dénotés ARIMA (p,d,q), traitent une série de valeurs ordonnées dans le temps et séparées par des intervalles équidistants comme un agencement linéaire soit des valeurs passées (processus autorégressif, q=0), soit de l'erreur aléatoire passée (processus de moyenne mobile, p=0) ou encore de la combinaison de ces deux composantes (processus mixte, p et q≠0), auquel s'ajoute une variation aléatoire résiduelle (ε_t). Le nombre de différenciation(s) préalable(s) à la stationnarité des séries est dénoté par «d» dans la notation conventionnelle. L'équation désignant le processus stochastique général (où y_t = série différenciée «d» fois) est la suivante:

$$y_t = \phi_1 y_{t-1} - \ldots - \phi_p y_{t-p} - \varepsilon_t - \theta_1 \varepsilon_{t-1} - \ldots - \theta_q \varepsilon_{t-q} \qquad (1)$$

3.2 Les fonctions de transfert

L'introduction d'une fonction de transfert permet de mettre en évidence, d'un point de vue statistique, la relation qui unit la fréquence cardiaque et le nombre de mouvements du bras dominant des travailleuses, en réponse au second objectif.

La formule générale mettant en relation la variable d'entrée (les bras) et la variable de sortie (la fréquence cardiaque) est de la forme suivante:

$$Y_t = \delta_1 Y_{t-1} - \ldots - \delta_r Y_{t-r} + \omega_0 X_{t-b} - \ldots - \omega_s X_{t-b-s} + N_t \qquad (2)$$

[inertie propre au rythme] [impact des bras] [bruit]

La fréquence cardiaque (Y_t) s'exprime ici par la somme de l'inertie propre au rythme, d'un présumé impact des bras (X_t) et d'une variation résiduelle indépendante des 2 autres composantes (N_t). A noter la présence d'un terme supplémentaire, «b», quantifiant le délai nécessaire à l'expression de l'effet de la variable d'entrée sur la variable de sortie (b pouvant être égale à zéro). En introduisant des opérateurs (B) dans l'équation 2, il est possible d'isoler le terme représentant la fonction de transfert V(B), c'est à dire l'équation permettant de passer directement de la variable «bras» à la variable «rythme» .

$$Y_t = \frac{\omega_0 + \omega_1 B + ... + \omega_s B^s}{1 - \delta_1 B - ... - \delta_r B^r} X_t + N_t \qquad (2.1)$$

3.3 Les modèles d'intervention

En regard du troisième objectif, on veut exploiter les variables opératoires qui rendent compte de façon plus précise de l'activité des travailleuses. On s'intéresse par exemple à comparer l'impact relatif des deux différents postes de travail occupés dans une journée c'est-à-dire l'*alimentation* et la *réception*[1] que l'on oppose aux temps d'arrêt. Le passage de la pause à l'un ou l'autre des 2 postes de travail peut être envisagé comme une intervention dans la série. On dispose donc de trois situations différentes pouvant être traduites en présence-absence par intervalle de temps sur l'échelle d'enregistrement de la fréquence cardiaque. Le modèle général intégrant les interventions est de la forme suivante:

$$Y_t = Z_t + X_t \qquad (3)$$

Le rythme cardiaque devient fonction d'une composante stochastique (X_t) telle que décrite par un modèle ARIMA(p,d,q), et d'une composante (Z_t) cumulant les interventions représentées par deux séries indicatrices:

$\xi_{a,t}$ = { 1 si à l'instant t la travailleuse est au poste de l'alimentation, 0 sinon
$\xi_{b,t}$ = { 1 si à l'instant t la travailleuse est au poste de la réception , 0 sinon

Le poste occupé par la travailleuse pouvant être entièrement prédit, les séries indicatrices sont de type déterministe et non stochastique. Le terme traduisant les interventions est dépourvu de variation résiduelle et se définit simplement par la somme des fonctions de transfert associées aux séries indicatrices:

$$Z_t = V_a(B)\,\xi_{a,t} + V_b(B)\,\xi_{b,t} \qquad (4)$$

4. Analyse de la variabilité et de la stabilité des réponses cardiaques

4.1 Considérations méthodologiques

L'étude de la variabilité, basée sur la comparaison des modèles obtenus à travers les trois approches modélisantes décrites, s'appuie sur deux niveaux descriptifs: 1- la forme du modèle ajusté; et 2- la valeur des paramètres du modèle.

L'examen de la forme et de l'ordre des modèles ajustés sur les différentes séries constitue le niveau le plus grossier d'analyse des différents jours-travailleuses. Cependant, on ne peut conclure que deux séries évoluent strictement suivant le même processus simplement parce que les modèles ajustés sont de même forme ou à l'inverse que deux séries diffèrent en tout point simplement parce que les modèles ajustés sont de formes différentes. Au-delà de la forme, la valeur des paramètres fournit une base quantitative de comparaison. Or, ce deuxième niveau de comparaison n'est possible que si les modèles sont de formes similaires.

Ceci nous renvoie aux problèmes de l'absence de modèle unique et à la subjectivité du choix du modèle. En l'absence de procédures automatisées bien implantées, l'étape d'identification du type de modèle reste essentiellement intuitive[6], et le problème de la subjectivité se pose avec beaucoup plus d'acuité, sachant que la démarche de modélisation ne conduit pas nécessairement à l'ajustement d'un modèle unique. Bien qu'en théorie on puisse s'appuyer sur le concept de modèle «optimal», en pratique, plusieurs modèles peuvent être ajustés et même validés sur une même série[7]. En conséquence, la comparaison de modèles, dans une perspective d'analyse de la variabilité, pourrait être fortement biaisée par l'ajustement «forcé» d'un modèle qu'on souhaiterait généralisé pour appuyer par exemple, l'hypothèse d'une grande stabilité de la réponse cardiaque.

Compte tenu de l'absence de modèle unique qui se surimpose à la subjectivité de l'identification du type de modèle, on a retenu une stratégie comportant 3 niveaux :

1- privilégier les modèles les plus simples: l'addition de paramètres supplémentaires n'étant retenue que pour valider le modèle;

2- opter pour la représentation dominante: si deux modèles aussi simples s'avèrent adéquats, la forme généralisée au plus grand nombre de séries étant choisie;

3- exprimer les modèles différents sous une forme commune.

Les deux premiers éléments de cette stratégie contribuent à systématiser la démarche de modélisation, alors que la réécriture des modèles permet de franchir le niveau de comparaison de la forme et d'examiner quantitativement le poids des interdépendances temporelles.

4.2 Réécriture des modèles

La procédure de réécriture des modèles s'applique tout aussi bien à la comparaison d'équations ajustées sur des séries distinctes qu'à la comparaison de modèles différents ajustés sur une même série.

La représentation de modèles différents sous une forme commune s'appuie sur le fait que les processus stochastiques considérés admettent une représentation en termes de bruit blanc: les processus de moyenne mobile étant déjà une sommation de variations aléatoires et les processus autorégressifs finis pouvant s'exprimer comme un processus de moyenne mobile infini[7].

Posons deux modèles différents, un processus de moyenne mobile et un mixte:

$$\text{a) ARIMA (0,1,3)} \quad y_t = \varepsilon_t - \alpha\varepsilon_{t-1} - \alpha_2\varepsilon_{t-2} - \alpha_3\varepsilon_{t-3} \tag{5}$$

$$\text{b) ARIMA (1,1,1)} \quad y_t = \varepsilon_t + \beta y_{t-1} - \alpha\varepsilon_{t-1} \tag{6}$$

En introduisant un opérateur (B) dans l'équation 5, on peut exprimer le modèle de moyenne mobile directement sous la forme d'un bruit blanc:

$$\text{a)} \quad y_t = \theta(B)\varepsilon_t \tag{7}$$

$$y_t = (1 - \alpha B - \alpha_2 B^2 - \alpha_3 B^3)\varepsilon_t \tag{7.1}$$

Dans le modèle mixte (équation 6), on introduit un opérateur à la composante autorégressive et un second opérateur à la composante de moyenne mobile:

$$\text{b)} \quad \varphi(B)y_t = \theta(B)\varepsilon_t \tag{8}$$

$$y_t = \varphi^{-1}(B)\,\theta(B)\varepsilon_t \tag{8.1}$$

$$\text{où } \theta(B) = (1 - \alpha B) \quad \text{et} \quad \varphi^{-1}(B) = \frac{1}{(1 - \beta B)}$$

Si dans l'équation 8.1, on exprime l'égalité en fonction de y_t, on obtient la notation inversée de l'opérateur φ. Pour ramener ce terme sous sa forme linéaire, on a recours aux séries géométriques de Taylor:

sachant que pour tout $|x| < 1$, $\frac{1}{1-x} = \sum_{i=0}^{\infty} x^i$

alors, $\frac{1}{(1-\beta B)} = 1 + \beta B + \beta^2 B^2 + \beta^3 B^3 + \ldots$

L'équation 8.1 peut s'écrire:

$$y_t = (1+ \beta B + \beta^2 B^2 + \beta^3 B^3 + \ldots)(1-\alpha B) \qquad (8.2)$$

$$y_t = \left(1-(\alpha-\beta)B - (\alpha\beta - \beta^2)B^2 - (\alpha\beta^2 - \beta^3)B^3 - \ldots \right)\varepsilon_t \qquad (8.3)$$

La forme réécrite du modèle mixte (équation 8.3) est maintenant comparable à celle du modèle de moyenne mobile (équation 7.1). Cette représentation commune de 2 modèles initialement différents, permet de comparer la valeur quantitative des pondérations appliquées aux opérateurs c'est-à-dire la valeur des interdépendances dans le temps.

4.3 Résultats

En exprimant des séries longues de plus de 3000 observations, par une équation paramétrique simple, comportant 3 ou 4 termes, on obtient une synthèse qui nous renseigne sur la structure de corrélation du rythme ou sur la relation entre celui-ci et des variables extrinsèques comme le mouvement des bras ou la nature du travail effectué. Pour illustrer la comparaison des modèles ajustés à travers les 3 différentes approches, on réfèrera aux 6 journées d'une même travailleuse.

4.3.1 Les séries chronologiques

La forme des modèles obtenus pour la modélisation du rythme cardiaque de la travailleuse 1 est relativement constante. La fréquence cardiaque des jours 1, 2, 3, 5 et 6 évolue suivant un processus de moyenne mobile d'ordre 3 alors que pour le jour 4, le modèle ajusté comporte une composante autorégressive et une composante de moyenne mobile.

A travers cette première étape d'analyse des modèles ajustés sur les séries de fréquence cardiaque, on peut mettre en évidence la stabilité relative de la structure de corrélation du rythme chez cette même personne, et détecter la variabilité, en l'occurrence, la composante autorégressive décrivant l'évolution du rythme au jour 4. Cette nuance dans l'évolution du rythme cardiaque sera explorée davantage: elle peut être fortuite et ne concrétiser qu'une part de la variabilité intra-individuelle mais elle peut également refléter l'influence de

facteurs extrinsèques ayant modifié, par exemple, la valeur moyenne du rythme ou la fréquence et l'amplitude des écarts à la moyenne.

Pour illustrer la comparaison des paramètres des modèles ajustés, on se limitera à la pondération associée au premier paramètre (délai = t-1) qui nous renseigne sur l'interdépendance des réponses cardiaques voisines. La valeur quantitative des pondérations est reportée en ordonnée sur le graphe de la figure 1. Le paramètre du modèle traduisant le rythme du jour 4 est extrait du modèle ARIMA (1,1,1) réécrit sous forme de bruit blanc (sans intervalle de confiance).

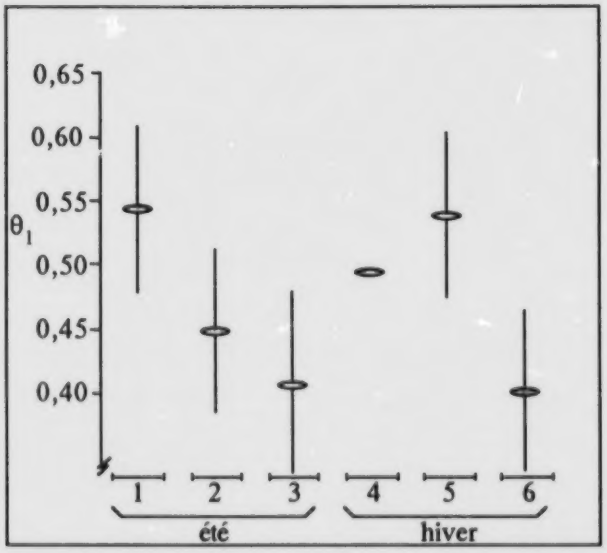


Fig 1: Valeurs et intervalles de confiance associés aux premiers paramètres des modèles ajustés sur les séries de fréquence cardiaque des 6 jours de la travailleuse 1.

Rappelons que le delai séparant chaque observation de fréquence cardiaque était de 30 secondes. Ainsi, les fluctuations de la fréquence cardiaque enregistrées sont nécessairement très fines et de fait, les pondérations attribuées au poids des valeurs voisines sont assez élevées, variant entre 0,40 et 0,65.

La plage de recoupement des intervalles de confiance est importante, ce qui encore une fois peut être interprété en termes de stabilité. Par ailleurs, la valeur moyenne des paramètres pour les jours d'hiver (4,5,6) semble légèrement plus élevée, laissant supposer une plus forte corrélation entre le rythme précédent ($_{t-1}$) et le rythme actuel ($_t$). Les hypothèses concernant le patron saisonnier ne seront vérifiables qu'avec l'ensemble des individus-jour. Si la tendance observée ici s'avère généralisée, on pourra postuler que l'hiver, en absence de surcharge cardiaque due à la lutte contre la chaleur, le rythme cardiaque est plus stable, l'accélération plus lente...

4.3.2 Les fonctions de transfert

L'identification des fonctions de transfert entre le nombre de mouvement(s) du bras dominant et la fréquence cardiaque n'a pu être effectuée que pour 5 des 6 jours du suivi de la travailleuse 1; l'enregistrement des mouvements des bras pour le jour 3 comportant trop de données manquantes. La comparaison de la forme des fonctions de transfert estimées et validées s'appuie sur l'équation permettant le passage direct de la série des bras vers la série de fréquence cardiaque (équation 2.1):

$$\text{jours 1, 2} \quad \frac{\omega_0}{1 - \delta_1 B} \qquad \text{jours 4, 5} \quad \frac{\omega_0 + \omega_1 B}{1 - \delta_1 B}$$

$$\text{jour 6} \quad \frac{\omega_0 + \omega_1 B}{1 - \delta_1 B - \delta_2 B^2 - \delta_3 B^3 - \delta_4 B^4}$$

Les fonctions de transfert des jours 1 et 2, correspondant aux jours d'été, sont de la forme relativement simple, soit une constante sur polynome de premier ordre. A l'opposé, on obtient des rapports entre des polynomes plus complexes pour les jours 4, 5 et 6, ce qui pourrait aller dans le sens de l'hypothèse formulée plus haut à savoir, une plus grande sensibilité de la fréquence cardiaque l'été, en présence de chaleur.

La quantification de l'impact des bras sur l'accélération cardiaque est beaucoup plus riche en informations que ne l'est l'expression mathématique de la relation entre ces deux variables. Après réécriture des fonctions de transfert, suivant une procédure similaire à celle décrite à la section 4.2, on peut reprendre l'équation linéaire mettant en relation les variables d'input et d'output (équation 2), et examiner la pondération associée à la variable «bras», en fonction des différents délais significatifs. La représentation de ces pondérations sous forme de gains permet de visualiser directement la réponse cardiaque dans le temps pour une augmentation de la vitesse des bras fixée à 1 mouvement (fig. 2).

L'influence de l'activité des bras sur l'accélération cardiaque semble peu soumise à la variabilité intra-individuelle. En effet, chez cette travailleuse, pour l'ensemble des jours échantillonnés, l'impact des bras est immédiat (délai 0 significatif), persiste au moins pendant les 3 premiers délais (90 secondes) et décroît de façon progressive. Les gains totaux, représentant l'augmentation nette de la fréquence cardiaque résultant de l'accroissement de la vitesse des bras d'un mouvement, sont sensiblement les mêmes et leurs valeurs élevées témoignent, en quelque sorte, de l'importance de l'effort cardiaque associé à la sollicitation des bras.

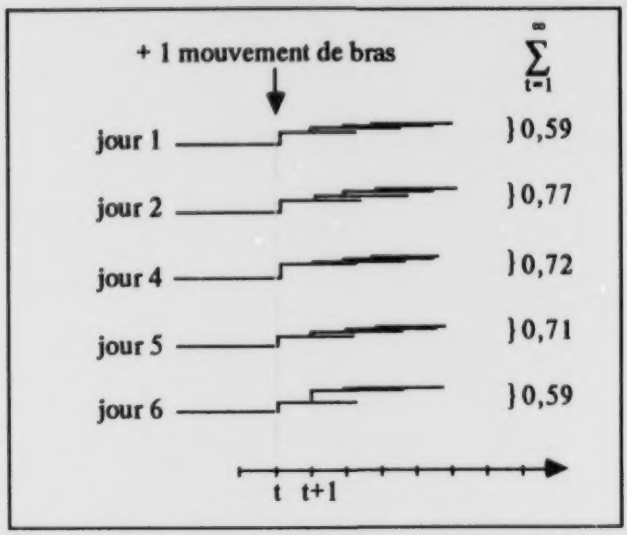


Fig. 2: Schématisation de la réponse cardiaque dans le temps pour une augmentation de la vitesse des bras d'un mouvement. Les gains totaux (sommation à l'infini des des gains partiels) sont également présentés pour 5 des 6 jours de la travailleuse 1.

4.3.3 Les modèles d'intervention

Les fonctions de transfert jumelées aux séries indicatrices représentant chacune des 2 interventions, «alimentation» et «réception» , sont de simples constantes et ce, pour les 6 jours du suivi de la travailleuse 1. L'intérêt, au-delà de cette forme commune, réside dans la possibilité de hiérarchiser l'impact relatif des 2 postes de travail qui impliquent des contraintes posturales spécifiques ainsi qu'un rythme de travail différent. Les constantes associées à l'impact respectif des 2 postes de travail sur la fréquence cardiaque, sont reportées à la figure 3.

On remarque que le passage de la pause à l'un ou l'autre des postes influence nettement la fréquence cardiaque, causant des augmentations variant de 4 à près de 15 battements/minute selon le jour et le poste. L'alimentation a un impact plus grand sur la fréquence cardiaque que la réception et ce, aussi bien en pairant les jours, qu'en imaginant les moyennes respectives. La valeur chiffrée des constantes associées à l'impact de chacun des postes varie et s'insère dans un large intervalle de confiance. Cette variabilité reflète en partie la sensibilité de la fréquence cardiaque et en partie la complexité des situations réelles de travail: la récupération en période de pause étant fonction de la pénibilité des activités précédentes et les phases de travail à l'alimentation ou à la réception n'étant pas nécessairement homogènes. Néanmoins, la constance et le sens des différences observées dans le comportement cardiaque en fonction du poste occupé, présentent un grand intérêt en vue d'identifier les phases de travail les plus contraignantes.

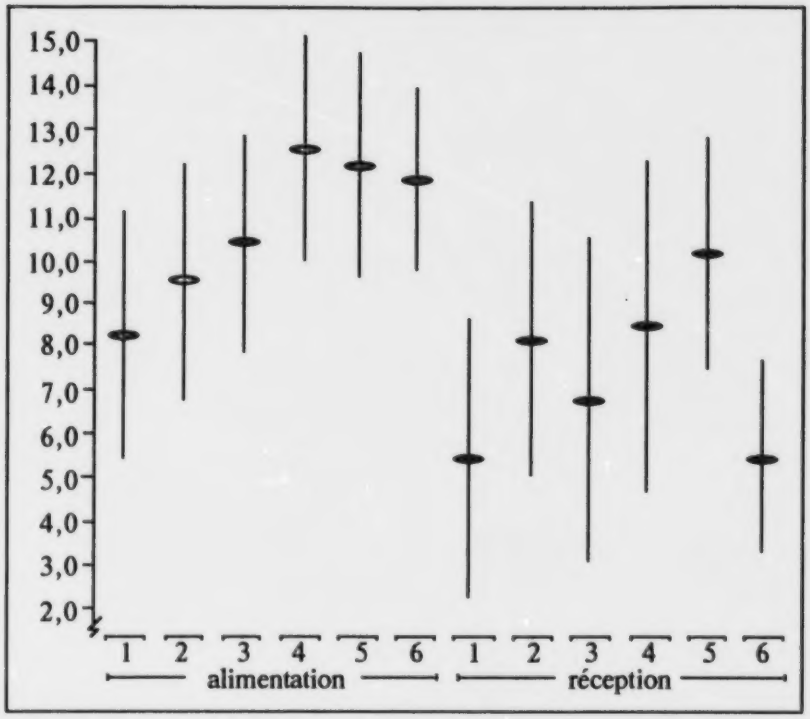


Fig. 3: Valeurs et intervalles de confiance des constantes associées à l'impact des 2 postes de travail pour les 6 jours de la travailleuse 1.

5. Conclusion et perspectives

Appliquée à l'ensemble des tracés de fréquence cardiaque dont nous disposons (66 jours-travailleuses) et raffinée pour y inclure d'autres variables explicatives, la démarche de modélisation nous permettra de mieux comprendre le comportement cardiaque en situation réelle de travail et plus spécifiquement, de mieux cerner les facteurs contibuant à l'accélération cardiaque à la base de l'astreinte observée. Dans cette perspective, la caractérisation de la variabilité/stabilité des réponses cardiaques, à travers l'analyse comparative des modèles ajustés, ajoute une portée méthodologique à l'intérêt théorique et pratique de l'application des techniques de modélisation chronologique aux phénomènes physiologiques.

Jusqu'ici essentiellement descriptive, l'analyse comparative des résultats de la modélisation sera orientée davantage vers une approche inférentielle. Nous envisageons des analyses de variance sur les paramètres des modèles, avec des facteurs incluant l'individu, le jour et la saison, mais également des co-variables puisées dans les autres volets de l'étude (température, âge, symptômes...). Parallèlement, nous misons sur le développement d'un test basé sur le principe du maximum de vraisemblance.

Sachant que la vraisemblance d'un groupe de paramètres est égale à la probabilité de réalisation d'une série et que les estimations des paramètres sont telles, qu'elles maximisent cette probabilité de réalisation, il est possible d'aborder la comparaison de modèles décrivant 2 séries en manipulant les maxima de vraisemblance.

En calculant le rapport (Λ) entre ce qui maximise le produit des probabilités de réalisation de 2 séries (à partir d'un même groupe de paramètres) et le produit des maxima de vraisemblance des 2 séries (estimées disctinctement), on obtient des valeurs variant entre 1 et 0, selon que les séries proviennent ou non d'un même processus.

soit $L(\beta) =$ Pr [réalisation de la série à partir des paramètres β]

$$\Lambda = \frac{\text{Max}_{\beta}\, L_1(\beta)\, L_2(\beta)}{\text{Max}_{\beta 1}\, L_1(\beta_1)\ \text{Max}_{\beta 2}\, L_2(\beta_2)}$$

Après transformation ($-2\log\Lambda$), le rapport calculé suit asymptotiquement une distribution du Chi carré. Une fois bien implanté, ce test permettra d'objectiver, d'un point de vue statistique, la différence ou l'équivalence de 2 ou plusieurs processus.

Références

1 Brabant, C.; Bédard, S.; Mergler, D. 1988. Cardiac strain among women laundry workers doing "light" work. Dans: Hogstedt, C.; Reuterwall, C. 1988. Progress in Occupational Epidemiology. Excerpta Medica International Congress series 829 (Amsterdam), pp. 237-241.

2 Frimat, P. 1988. La fréquence cardiaque, variations, limites d'interprétation. Actes du colloque national "Fréquence cardiaque et travail" (Lille, novembre 1987). Revue de Médecine du Travail, XV(4): 829-830.

3 Brabant, C.; Ferraris, J.; Latour, A.; Mergler, D. 1988. Application de différentes techniques de modélisation à l'étude de l'évolution de la fréquence cardiaque. Recueil des textes des présentations du colloque sur les méthodes et domaines d'application de la statistique. (Moncton, mai 1988). Bureau de la statistique du Québec (Québec), pp. 293-306.

4 Box, G.E.P.; Jenkins, G.M. 1976. Times series analysis: Forecasting and Control. Holden-Day (San Francisco), 575 pages.

5 Box, G.E.P.; Tiao, G.C. 1975. Intervention analysis with applications to economic and environmental problems. J Amer Statist Assoc, 70:70-79.

6 Raftery, A.E. 1985. Times series analysis. Eur J Operational Research, 20:127-137.

7 Chatfield, C. 1984. The analysis of time series: an introduction. Troisième édition. Chapman and Hall (London), 286 pages.

Application d'estimateurs robustes pour le traitement de séries chronologiques environnementales.

Claude LABERGE
INRS–Eau, Université du Québec, 2800 Einstein
C.P. 7500 Sainte Foy, P.Q., G1V 4C7.

1. Introduction

Les caractéristiques particulières des données disponibles dans le domaine de l'eau ont amené plusieurs auteurs à développer ou à adapter des méthodes statistiques qui permettent une analyse efficace de ces ensembles de données. Dans les études de détection de tendance, les méthodes non paramétriques permettent de réduire l'influence des valeurs aberrantes et des distributions non normales mais des adaptations demeuraient nécessaires afin de tenir compte de l'autocorrélation (LETTENMAIER 1976), des saisonnalités (HIRSCH et al 1982) ainsi que de la présence simultanée de ces deux types de dépendances à court terme (HIRSCH et SLACK 1984). Un désavantage important des méthodes non paramétriques pour les environnementalistes est l'absence de connaissance du niveau de la série et de l'amplitude de la tendance. La présente communication se veut une introduction vers l'utilisation des méthodes robustes de HUBER pour la détection des tendances dans le domaine de l'eau. La section 2 situera le cadre théorique dans lequel les applications seront développées. La section 3 présentera deux applications sur des séries de qualité et de quantité des eaux du Lac Laflamme, P.Q. A la suite de ces résultats, on discutera de la possibilité d'utilisation des méthodes robustes dans le cadre d'études de détection de tendances.

2. Cadre théorique

Les applications qui seront présentées à la section suivante reposent sur les développements théoriques contenus dans LABERGE 1988, le lecteur devra donc s'y référer pour une description plus approfondie des méthodes utilisées. On suppose ici que les valeurs

aberrantes présentes dans les séries étudiées sont de type statique (VAS) ou de type dynamique (VAD), correspondant respectivement aux termes anglais "Additive Outlier" et "Innovation Outlier" utilisés par FOX 1972. On distiguera alors entre trois processus: 1) le processus sous–jacent que l'on ne peut mesurer mais dont on veut estimer les paramètres; 2) le processus de valeurs aberrantes qui perturbent les mesures recueillies mais que l'on ne peut mesurer; 3) le processus mesuré qui est une combinaison des deux processus mentionnés précédemment. Le but de l'utilisation des méthodes robustes est alors d'utiliser les observations du processus mesuré pour estimer les paramètres du processus sous–jacent sans être affecté par la présence des valeurs du processus de valeurs aberrantes. Il est bon de noter que les estimateurs classiques (Estimateur des Moindres Carrés (EMC) et Estimateurs du Maximum de Vraisemblance Conditionnel (EMVC)) estiment les paramètres du processus mesuré. Les estimateurs robustes utilisés dans les applications de la section 3 sont les M–Estimateurs et les GM–Estimateurs avec la fonction de HUBER.

L'étude de l'estimation ponctuelle des paramètres μ, moyenne du processus, ϕ, le paramètre d'autorégression d'ordre 1, et σ, l'écart type des impulsions, permet les conclusions suivantes (LABERGE 1988):

> Pour les processus gaussiens, tous les estimateurs convergent vers les paramètres du processus sous–jacent qui est alors équivalent au processus mesuré puisqu'il y a absence de valeurs aberrantes. Les estimateurs robustes sont cependant moins efficaces que les EMVC, mais la perte d'efficacité, qui dépend du choix de la constante de winsorisation, peut être contrôlée.
>
> Pour les processus avec VAD–0 (les valeurs aberrantes dynamiques ont une moyenne de 0, tout comme le processus sous–jacent), tous les estimateurs de μ et ϕ convergent vers les estimateurs du processus sous–jacent. Les estimateurs robustes sont cependant plus efficaces pour l'estimation de ces deux paramètres. En ce qui concerne l'estimation de σ, les estimateurs robustes convergent vers une quantité près du paramètre d'échelle du processus sous–jacent, ce qui n'est pas le cas pour l'EMVC.
>
> Pour les processus avec VAS–0 (les valeurs aberrantes statiques ont une moyenne de 0, tout comme le processus sous–jacent), tous les estimateurs de μ convergent vers le paramètre de localisation du processus sous–jacent et semblent aussi efficaces les uns que les autres. Pour l'estimation de ϕ, tous les estimateurs sont biaisés, mais

les GM–Estimateurs présentent les biais les plus faibles. Finalement, en comparaison avec l'EMVC, les estimateurs robustes du paramètre d'échelle convergent vers une quantité beaucoup plus près du paramètre d'échelle du processus sous–jacent.

Pour les processus avec VAD–10 (les valeurs aberrantes dynamiques ont une moyenne de 10, alors que le processus sous–jacent a une moyenne de 0), tous les estimateurs de ϕ convergent vers le paramètre associé au processus sous–jacent. En ce qui concerne l'estimation de μ et de σ, les estimateurs robustes convergent vers des quantités près des paramètres de localisation et d'échelle du processus sous–jacent, ce qui n'est pas le cas pour les EMVC.

Finalement, pour les processus avec VAS–10 (les valeurs aberrantes statiques ont une moyenne de 10, alors que le processus sous–jacent a une moyenne de 0), aucun estimateur de σ et ϕ ne converge vers les paramètres de processus sous–jacent. Pour l'estimation de ϕ, les GM–Estimateurs présentent cependant un avantage sur les autres estimateurs, puisque leur biais est beaucoup plus faible. Pour l'estimation de μ, les GM–Estimateurs convergent vers une quantité très proche du paramètre de localisation du processus sous–jacent.

L'estimation de μ par intervalle de confiance utilise des estimateurs de μ, ϕ et σ, les résultats des paragraphes précédents s'appliquent donc à cette estimation. On devrait donc donner plus d'importance à l'estimation classique (EMVC) lorsque aucune valeur aberrante n'est présente. Lorsque des VAD sont présentent, on devrait plutôt donner un certain avantage aux M–Estimateurs à cause de leur plus grande efficaité. Finalement, la présence de VAS devrait mener l'utilisateur vers les GM–Estimateurs qui sont les moins biaisés pour l'estimation des paramètres du processus sous–jacent.

3. Applications

Les séries chronologiques étudiées ont été échantillonnées au Lac Laflamme dans la forêt Montmorency à environ 75 km au nord–est de la ville de Québec. Ces séries sont constituées de mesures de concentrations (mg/l) au centre et à l'exutoire (sortie) du lac ainsi que des mesures de charges (kg/hectare.semaine) à l'exutoire. Les mesures de plusieurs ions sont disponibles dans l'ensemble de données mais on s'attardera seulement au

paramètre fer pour les applications présentées ici. Les séries contiennent 300 observations hebdomadaires débutant le 16 mars 1981 et se terminant le 9 décembre 1986. La première application étudiera les concentrations de fer au centre et à l'exutoire du lac, on cherchera alors à voir les différences physiques entre les deux sites à l'aide de l'estimation des différents paramètres. La deuxième application s'attardera aux charges à l'exutoire, on illustrera alors l'intérêt de l'utilisation des estimateurs robustes en présence de deux sous–populations distinctes.

3.1 Etude des concentrations.

La figure 3.1a présente la série de concentrations de fer au centre du lac alors que la figure 3.1b présente la série de concentrations de fer à l'exutoire du lac. On peut noter la présence de plusieurs valeurs douteuses qui peuvent influencer les différents estimateurs. Dans les deux cas, l'identification d'un modèle AR(1) est valable; l'estimation d'un tel modèle est donc d'intérêt. Les figures 3.1c et 3.1d présentent respectivement les histogrammes des concentrations des deux séries, il est alors possible de remarquer que les valeurs douteuses se retrouvent toutes du côté droit de la distribution. Ces figures proposent donc la présence de valeurs aberrantes de moyenne différente de la moyenne du processus sous–jacent. Les figures 3.1e et 3.1f illustrent finalement les histogrammes des résidus du modèle AR(1) estimé à l'aide des M–Estimateurs de HUBER. Ces deux dernières figures permettent de constater que malgré la présence de valeurs aberrantes uniquement du côté droit de la distribution, certains résidus apparaissent aberrants des deux côtés de la distribution des résidus. Cette dernière affirmation propose la présence de VAS qui sont caractérisées par la présence de deux résidus aberrants successifs de signes inverses.

Figure 3.1 Séries de concentrations de fer (mg/l) au centre et à l'exutoire du lac Laflamme et histogrammes de fréquence des observations et des résidus (M–Estimateurs, modèle AR(1)).

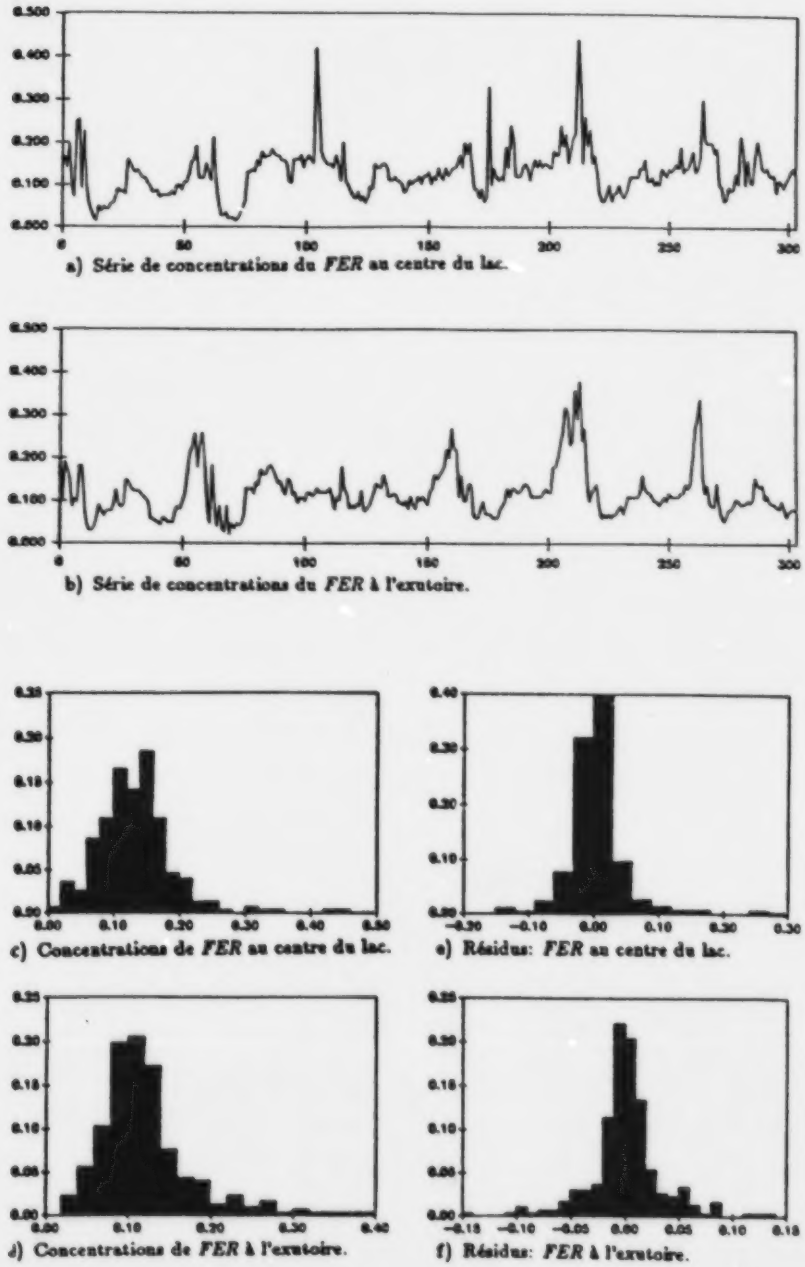


Tableau 3.1. Estimateurs robustes et classiques pour les paramètres de localisation, d'échelle et d'autorégression; séries de concentrations de fer au centre du lac et à l'exutoire.

LAC

Estimateur	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\phi}$	I.C.(μ)
EMVC	0.13	0.07	0.64	[.11 .15]
M–Estimateur	0.12	0.02	0.74	[.10 .13]
GM–Estimateur	0.11	0.02	0.82	[.10 .13]

EXUTOIRE

Estimateur	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\phi}$	I.C.(μ)
EMVC	0.12	0.04	0.82	[.09 .14]
M–Estimateur	0.11	0.02	0.86	[.10 .13]
GM–Estimateur	0.11	0.02	0.85	[.10 .13]

Le tableau 3.1 présente les résultats des estimations classiques et robustes pour les trois paramètres associés au modèle AR(1). L'utilisation des estimateurs du maximum de vraisemblance amène à la conclusion que l'autocorrélation à l'exutoire (.82) est significativement plus grande que l'autocorrélation au centre du lac (.64). Cependant, les GM–Estimateurs n'appuient pas cette observation puisque les estimateurs de ϕ sont respectivement de .82 au centre du lac et .85 à l'exutoire. La présence de valeurs aberrantes statiques plus importantes dans la série au centre du lac est sans aucun doute la cause de ces résultats différents. Le M–Estimateur au centre du lac appuie cette possibilité puisqu'il présente un résultat intermédiaire, ce M–Estimateur de ϕ (.74) amène une gradation caractéristique (EMVC de ϕ < M–E de ϕ < GM–E de ϕ) des estimateurs de ϕ en présence de VAS. En ce qui concerne la série à l'exutoire, les estimateurs permettent de conclure que l'influence de VAS est faible. Devant de telles observations, les GM–Estimateurs apparaissent les plus adéquats pour l'estimation de la série au centre du lac alors que pour la série à l'exutoire, les trois estimateurs sont sensiblement équivalents. Il faut alors conclure que l'autocorrélation n'est pas significativement différente aux deux sites du lac, la différence au niveau des EMVC de ϕ étant introduite par la présence de

quelques valeurs aberrantes statiques.

Les estimateurs robustes de μ et σ permettent également de conclure qu'il n'existe pas de différence significative entre le centre du lac et l'exutoire. Il semble donc que les deux séries présentent des carcatéristiques semblables, il faut cependant noter que les valeurs aberrantes aparaissent plus importantes dans la série au centre du lac ce qui amène des résultats significativement différents avec l'utilisation des estimateurs classiques.

3.2 Etude des charges

La figure 3.2 présente la série des charges de fer exportées à l'exutoire du lac Laflamme.

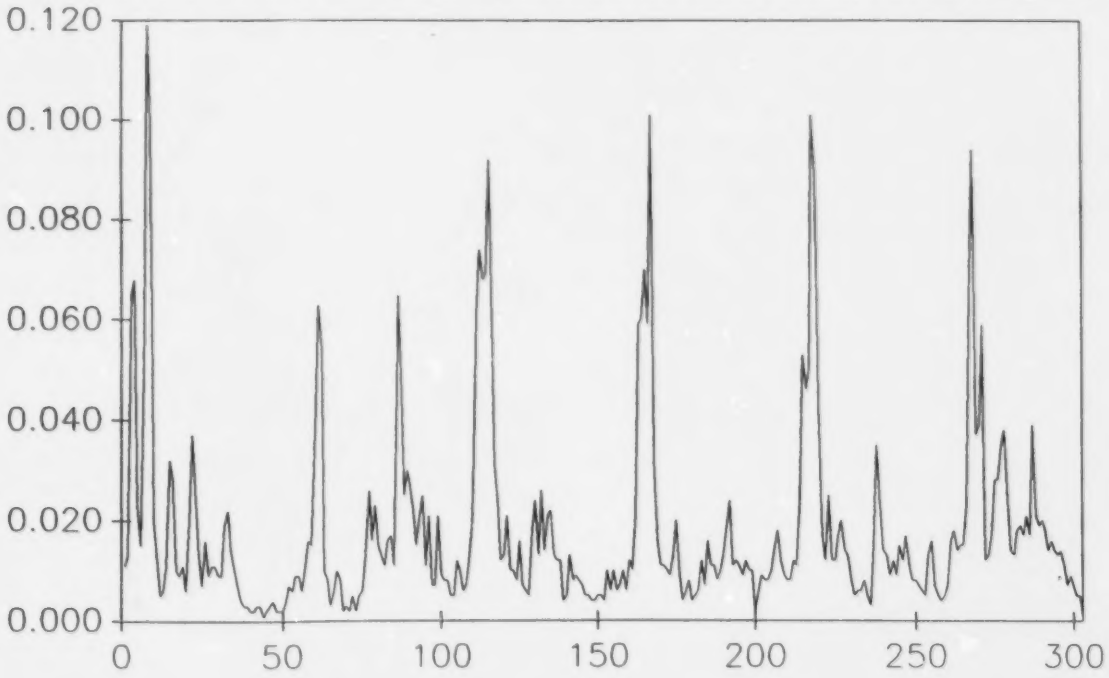


Figure 3.2 Séries de charges de fer (kg/hectare.semaine) exportées à l'exutoire du lac Laflamme.

On remarque la présence évidente d'une saisonnalité: apparition à toutes les 52 semaines d'une forte charge exportée. Cette saisonnalité est introduite par les grandes crues de printemps qui peuvent alors être vues comme une sous–population différente du reste de l'année où les débits sont généralement normaux (certains débits peuvent être élevés à

l'automne). Etant donné que les crues de printemps ne constituent que quelques semaines sur l'année complète, on peut être intéressé à estimer les caractéristiques de la série sans être affecté par les valeurs associées au crues de printemps. Dans une telle situation, les estimateurs robustes permettent une estimation rapide et simple: la série des charges constituent alors la série mesurée, les valeurs associées au crues de printemps constituent la série des valeurs aberrantes et les valeurs associées à des débits normaux constituent le processus sous–jacent dont on veut estimer les paramètres.

Tableau 3.2. Estimateurs robustes et classiques pour les paramètres de localisation, d'échelle et d'autorégression; série de charges de fer à l'exutoire.

EXUTOIRE

Estimateur	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\phi}$	I.C.(μ)
EMVC	0.018	0.019	0.72	[.012 .024]
M–Estimateur	0.011	0.006	0.77	[.009 .015]
GM–Estimateur	0.011	0.006	0.74	[.009 .014]

Le tableau 3.2 présente les résultats des estimations classiques et robustes pour la série de charges exportées à l'exutoire. Les estimateurs de ϕ très semblables (non significativement différents) permettent de croire que les valeurs aberrantes sont surtout du type dynamique, ce qui est en accord avec l'idée que les crues printanières sont introduites dans le système et affectent les valeurs futures. Les estimateurs de ϕ montrent donc que la corrélation entre les données successives est assez élevée, soit d'environ 0.7.

Les estimateurs robustes de μ (0.011 kg/hectare.semaine) sont significativement différents de l'estimateur classique (0.018 kg/hectare.semaine). Les valeurs aberrantes apparaissant d'un seul côté de la distribution amène cette différence significative. On peut alors conclure que les estimateurs robustes nous donnent une estimation des charges exportées par débits normaux (environ 90% du temps) alors que l'estimateur classique nous donne un estimateur du niveau moyen de la série mesurée. Cette dernière estimation devrait être utilisée pour estimer la charge totale exportée durant l'année, mais elle serait inadéquate pour estimer la charge exportée à une semaine donnée puisqu'elle donnerait une estimation trop petite en temps de crue printanière et une estimation trop grande en temps de débit

normal. L'estimation par intervalle de confiance pour μ, illustre de nouveau la différence entre les estimateurs robustes et l'EMVC, alors que les intervalles robustes et classique n'ont qu'une faible partie commune.

Une décomposition artificielle de la série mesurée montre que la moyenne des observations plus petite que .025 (83% des semaines échantillonnées) est de .011 (écart type de .006) alors que la moyenne des observations supérieures à .025 est de .053 (écart type de .024). Ces mesures peuvent alors être vues comme les moyennes associées respectivement aux débits normaux et aux crues printanières (parfois automnales). Ces résultats illustrent de nouveau que les estimateurs classiques sous–estimeront la charge exportée en temps de crue et surestimeront en temps de débits normaux. Il apparaît donc d'intérêt d'utiliser conjointement les estimateurs robustes et classiques afin d'obtenir le plus d'information possible sur les caractéristiques de la série.

4. Conclusion

Les applications présentées ont permis de mettre en évidence l'intérêt de l'utilisation des estimateurs robustes pour le traitement des séries chronologiques disponibles dans le domaine de l'eau. L'utilisation conjointe avec les estimateurs classiques permet de faire ressortir plusieurs caractéristiques des séries étudiées: présence ou non de valeurs aberrantes, type des valeurs aberrantes, amplitude des valeurs aberrantes, etc. Devant de tels résultats il apparaît plausible de croire que les estimateurs robustes seraient d'une grande utilité pour la détection des tendances dans le domaine de l'eau. Les limites seraient cependant le peu de disponibilité de support informatique et le caractère asymptotique des tests robustes. Leur grand avantage serait la possibilité de connaître le niveau de la série et l'amplitude de la tendance, sans que les estimations soient influencées par la présence de valeurs aberrantes, ce qui n'est pas possible avec l'utilisation des tests non paramétriques pour la détection des tendances.

5. Bibliographie

Fox,A.J.: 1972, 'Outliers in Time Series' JRSSS, Ser B. 34, 350–363.

Hirsch,R.M., Slack,J.R., et Smith,R.A.: 1982, 'Techniques of Trend Analysis for Monthly Water Quality Data' Water Resources Research, vol. 18,107–121.

Hirsch,R.M., et Slack,J.R.: 1984, 'A Non–Parametric Trend Test for Seasonal Data with Serial Dependence' Water Resources Research, vol. 20,727–732.

Laberge,C.: 1988, 'Estimation Robuste des Paramètres de modèle AR(1): Application à des données Environnementales' Mémoire de maîtrise, Université Laval.

Lettenmaier,D.P.: 1976, 'Detection of Trends in Water Quality Data from Records with Dependents Observations' Water Resources Research, vol. 12,1037–1046.

Le choix d'une méthode de prévision

ALAIN LATOUR
Département de mathématiques et informatique,
Université du Québec à Montréal.

1 INTRODUCTION

Lorsque se pose le problème de la prévision, apparaît de façon naturelle le problème du choix de la méthode de prévision. Un gestionnaire peut désirer prévoir le comportement d'une variable qui l'intéresse dans le but de réduire l'incertitude qu'il a face à celle-ci. La connaissance immédiate de la valeur "probable" de cette variable peut aider le gestionnaire à prendre aujourd'hui des décisions qui lui seront favorables demain. Nous disposons maintenant de plusieurs outils de prévision: modèles économétriques, modèles de régression, méthode de Box et Jenkins, lissage exponentiel, etc. ...

Dans cet article nous examinerons principalement trois approches de prévision. La première, essentiellement basée sur la régression, propose des modèles du type:

$$Z_t = \varphi(t) + \varepsilon_t \qquad (1)$$

où $\varphi(t)$ est une composante déterministe tandis que ε_t est une erreur aléatoire de moyenne nulle. Dans bien des cas on suppose que $\varphi(t) = \alpha_0 + \alpha_1 t + \ldots + \alpha_k t^k$, avec k = 1, 2 ou 3. On peut alors envisager l'estimation des paramètres par la méthode des moindres carrés. Lorsque les données affichent une tendance nette, une telle approche peut s'avérer très satisfaisante. Lorsque des composantes cycliques sont nécessaires, on peut y introduire des composantes sinusoïdales.

Dans certaines situations, les méthodes de lissage peuvent être préférables. Elles permettent un ajustement dynamique des paramètres du modèle et sont d'une grande simplicité au niveau de l'implémentation sur ordinateur.

Finalement, le gestionnaire peut aussi envisager l'application de la méthodologie de Box et Jenkins. Cependant, cette approche nécessite une bonne connaissance des propriétés des modèles ARIMA afin de pouvoir faire une identification juste et adéquate du modèle de prévision. Au niveau des calculs impliqués, cette approche requiert un support informatique considérable, tout au moins au niveau de l'estimation des paramètres du modèle qui doit être faite via une procédure d'optimisation non linéaire.

Afin de démontrer l'intérêt de diverses méthodes, nous les appliquerons à la série bien connue du nombre de passagers des lignes aériennes pour la période allant de janvier 1948 à décembre 1960. Les 12 derniers mois feront l'objet d'une prévision afin de comparer les différentes méthodes.

Nous prenons pour acquis que le lecteur est familier avec les techniques de régression. Nous ne nous attarderons pas sur le modèle ARIMA suggéré par BOX et JENKINS (1976). Nous nous contenterons d'en donner l'équation et nous l'utiliserons pour faire des prévisions. Notre attention sera plutôt portée sur les méthodes de lissage.

2 LE LISSAGE EXPONENTIEL

Afin d'introduire les méthodes de lissage, considérons le modèle simple

$$Z_t = \mu + \varepsilon_t, \quad \text{où } \varepsilon_t\text{: i.i.d. } (0,\sigma^2) \qquad (2)$$

Nous dénoterons $Z_t(k)$ la valeur prédite au temps t de Z_{t+k}. Si le paramètre μ est connu, il est raisonnable d'utiliser comme prévision

$$Z_t(k) = \mu,\ k \geq 1.$$

L'erreur quadratique moyenne de cette prévision est

$$EQM = E\{(Z_{t+k} - \mu)^2\} = \sigma^2, \qquad (3)$$

Dans un contexte un peu plus réaliste, μ n'est pas connu. On peut alors prendre:

$$Z_t(k) = \bar{X} = \frac{1}{t}\sum_{i=1}^{t} Z_i,\ k \geq 1. \qquad (4)$$

Il s'agit d'une prévision sans biais dont l'EQM est $\left(1 + \frac{1}{n}\right)\sigma^2$. Dans ce contexte, on peut démontrer que les prévisions peuvent être mises à jour via la récurrence

$$Z_{n+1}(1) = Z_n(1) + \left(1 - \frac{1}{n}\right)(Z_{n+1} - Z_n(1)) \qquad (5)$$

Cette approche prévisionnelle peut se justifier dans la mesure où le paramètre μ est très stable. Dans certains cas où cette hypothèse n'est pas vérifiée, il se peut que μ soit "localement" constant, c'est-à-dire que μ varie peu et non systématiquement. On suggère alors de faire un prévision accordant plus d'importance aux données les plus récentes. On envisage alors un estimateur de la forme

$$Z_n(k) = c\sum_{t=0}^{n-1} \omega^t Z_{n-t} \qquad (6)$$

où

$$\omega \in (0, 1) \text{ et } c = \frac{1 - \omega^n}{1 - \omega},$$

nous assurant ainsi d'un ensemble de poids dont le total est 1. On peut alors montrer que

$$Z_{n+1}(1) = Z_n(1) + (1 - \omega)(Z_{n+1} - Z_n(1)) \qquad (7)$$

Si dans (5) on pose $\omega = \frac{n}{n+1}$, on retrouve (4). On préfère parfois écrire (5) sous la forme

$$Z_{n+1}(1) = \alpha Z_{n+1} + (1 - \alpha)Z_n(1) \qquad (8)$$

où $\alpha = 1 - \omega$, ce qui signifie que la valeur prise par la variable Z au temps n + 2 est prédite par une combinaison linéaire de la valeur prise au temps n + 1 et de la prévision faite au temps n de la valeur prise au temps n + 1.

Si nous ne pouvons supposer que la moyenne est "localement constante", il se peut qu'elle soit "localement linéaire". Pour traduire l'idée que la moyenne varie de façon linéaire, nous pouvons écrire

$$\mu_{n+1} = \mu_n + \beta \qquad (9)$$

Au temps n + 1, deux estimations de μ_{n+1} sont disponibles: Z_{n+1}, dont l'espérance est μ_{n+1}, et $\hat{\mu}_n + \beta$. Cette dernière estimation n'est réalisable que si β est un paramètre connu. Dans le cas où β est inconnu, on peut envisager une estimation dynamique de celui-ci. Soit $\hat{\beta}_n$, l'estimation du paramètre β faite au temps n. On prendra

$$\hat{\mu}_{n+1} = \alpha_1 Z_{n+1} + (1 - \alpha_1)(\hat{\mu}_n + \hat{\beta}_n) \qquad (10)$$

où $\alpha_1 \in (0, 1)$ est une constante. De plus, au temps n + 1, le paramètre β peut être estimé d'une part, par $\hat{\mu}_{n+1} - \hat{\mu}_n$, et d'autre part, par $\hat{\beta}_n$. Ce qui nous conduit à la récurrence

$$\hat{\beta}_{n+1} = \alpha_2(\hat{\mu}_{n+1} - \hat{\mu}_n) + (1 - \alpha_2)\hat{\beta}_n. \qquad (11)$$

Les prévisions sont alors données par

$$Z_{n+1}(k) = \hat{\mu}_{n+1} + \hat{\beta}_{n+1}k \qquad (12)$$

On trouve un excellent exposé sur les méthodes de lissage dans ABRAHAM et LEDOLTER (1983).

3 DIFFÉRENTES APPROCHES POUR UN MEME PROBLEME

Une série bien connue et grandement exploitée est celle du nombre de passagers des lignes aériennes (série G, BOX et JENKINS, 1976). Il s'agit d'une donnée mensuelle pour la période allant de janvier 1948 à décembre 1960. Avec différents modèles, nous essayerons de prévoir les valeurs de la dernière année. Par conséquent, le choix des paramètres des modèles envisagés, voire l'estimation de ceux-ci, sera basé uniquement sur les données allant de janvier 1948 à décembre 1959. Le graphique de ces données est présenté à la figure 1. Comme on peut le constater, la variance de cette série chronologique semble augmenter avec t. Il peut donc s'avérer préférable de travailler avec le logarithme de la variable, comme en témoigne la figure 2.

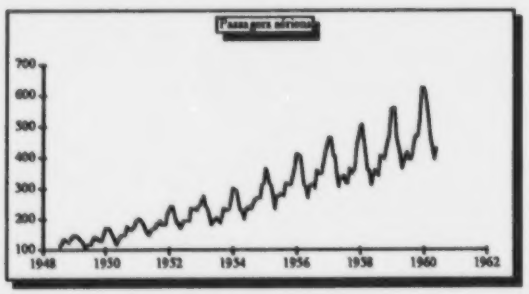


Figure 1: Nombre de passagers des lignes aériennes

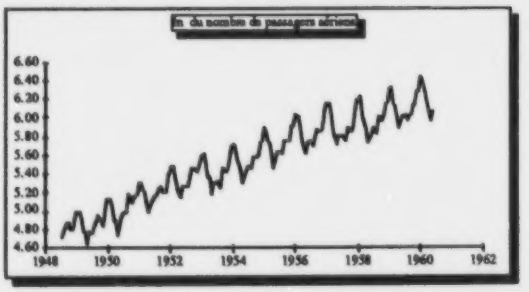


Figure 2: Logarithme du nombre de passagers des lignes aériennes

3.1 RÉGRESSION AVEC VARIABLES INDICATRICES

Le premier modèle que nous suggérons afin de représenter ces données est

$$Z_t = \alpha_0 t + \sum_{i=1}^{12} \alpha_i \delta_{t,i} + \varepsilon_t \qquad (13)$$

où

$$\delta_{t,i} = \begin{cases} 1 & t \equiv_{12} i \\ 0 & \text{autrement} \end{cases},$$

où "$t \equiv_{12} i$" signifie que t est congru à i, modulo 12. Ainsi, les paramètres α_i, $1 \le i \le 12$ sont des ajustements mensuels faits à la valeur $\alpha_0 t$. En appliquant un programme standard de régression on obtient un modèle ayant un coefficient d'explication $R^2 = 99.99\%$. Les estimations produites par la procédure GLM de SAS sont données au tableau 1. Au tableau 2, nous présentons l'analyse de la variance pour ce modèle.

Variables	d.l.	Paramètres estimés	Erreur standard	Valeur t	Prob. > t
t	1	0.010	.0001	78.302	0.0001
$\delta_{t,1}$	1	4.715	0.019	247.172	0.0001
$\delta_{t,2}$	1	4.698	0.019	245.535	0.0001
$\delta_{t,3}$	1	4.834	0.019	251.923	0.0001
$\delta_{t,4}$	1	4.792	0.019	248.985	0.0001
$\delta_{t,5}$	1	4.788	0.019	248.027	0.0001
$\delta_{t,6}$	1	4.910	0.019	253.604	0.0001
$\delta_{t,7}$	1	5.011	0.019	257.996	0.0001
$\delta_{t,8}$	1	5.004	0.019	256.835	0.0001
$\delta_{t,9}$	1	4.863	0.020	248.818	0.0001
$\delta_{t,10}$	1	4.721	0.020	240.834	0.0001
$\delta_{t,11}$	1	4.581	0.020	232.907	0.0001
$\delta_{t,12}$	1	4.696	0.020	238.024	0.0001

Tableau 1.

Source	d.l.	Somme des carrés	Moyenne quadratique	Valeur F	Prob. > F
Modèle	13	3995.777	307.367	93316.7	0.0001
Erreur	119	0.392	0.003		
Total	132	3996.169			
	R^2=0.999				

Tableau 2.

3.2 Régression harmonique

Il peut sembler tout à fait raisonnable de tenter de prévoir le développement de cette série en utilisant un modèle de la forme

$$Z_t = \beta_0 + \beta_1 t + \sum_{i=1}^{12} \beta_{1,i} \sin\varphi_i t + \beta_{2,i} \cos\varphi_i t + \varepsilon_t \qquad (13)$$

Les estimations des paramètres sont présentées au tableau 3 et l'analyse de la variance est donnée au tableau 4.

Variables	d.l.	Paramètres estimés	Erreur standard	Valeur t	Prob. > t
β_0	1	4.801	0.010	477.865	0.0001
β_1	1	0.010	.0001	78.560	0.0001
$\beta_{1,1}$	1	-0.046	0.017	-6.533	0.0001
$\beta_{2,1}$	1	-0.140	0.017	-19.835	0.0001
$\beta_{1,2}$	1	0.079	0.017	11.174	0.0001
$\beta_{2,2}$	1	-0.025	0.017	-3.557	0.0005
$\beta_{1,3}$	1	-0.001	0.017	-1.413	0.1603
$\beta_{2,3}$	1	0.027	0.017	3.840	0.0002
$\beta_{1,4}$	1	0.024	0.017	3.476	0.0007
$\beta_{2,4}$	1	0.025	0.017	3.519	0.0006
$\beta_{1,5}$	1	0.022	0.017	3.107	0.0024
$\beta_{2,5}$	1	0.006	0.017	0.781	0.4363

Tableau 3.

Source	d.l.	Somme des carrés	Moyenne quadratique	Valeur F	Prob. > F
Modèle	13	22.3018	2.027	619.385	0.0001
Erreur	119	0.3927	0.003		
Total	132	22.6946			
	R^2=0.983				

Tableau 4.

3.3 Modèle additif et modèle multiplicatif de Winters

3.3.1 MODELE ADDITIF

Le modèle additif de Winters suppose que

$$Z_n = T_n + S_n + \varepsilon_n \qquad (14)$$

où $T_{n+j} = \mu_{n+j} = \mu_n + \beta j$ est une tendance linéaire, tandis que $\{S_n\}$ est une suite de constantes telles que $S_n = S_{n+ks}$, $\forall\, k \in \{0, \pm1, \pm2, \ldots, \}$ et $\forall\, n \in \{1, \ldots, s\}$. Ces constantes sont des ajustements devant être faits aux valeurs T_n afin de tenir compte du caractère cyclique de la série. On suppose de plus que $\sum_{i=1}^{s} S_i = 0$. Il s'agit donc de faire une estimation dynamique des paramètres μ_n, S_n et β. Tout comme à la section 2, au temps n+1, comme estimation de μ_{n+1} nous disposons de la valeur de Z_{n+1}, qui doit cependant être ajustée par S_{n+1}. D'une part, nous savons que $S_{n+1} = S_{n-s+1}$. D'autre part, la suite des constantes S_i étant inconnue, nous pouvons envisager d'estimer dynamiquement ces termes, S_i, $1 \le i \le s$. Soit $\hat{S}_n$ l'estimation faite de S_n au temps n. Pour estimer μ_{n+1} au temps n + 1 nous pouvons donc considérer

$$\hat{\mu}_{n+1} = Z_{n+1} - \hat{S}_{n+1} \qquad (15)$$

Cependant, $\hat{\mu}_n + \hat{\beta}_n$ est aussi un estimateur de μ_{n+1}. Il faut noter ici que β_n est un paramètre inconnu que nous devons estimer. Cette estimation, au temps n + 1 est dénotée par $\hat{\beta}_{n+1}$. On utilise alors les deux estimations de μ_{n+1} via une combinaison linéaire des deux:

$$\hat{\mu}_{n+1} = \alpha_1(Z_{n+1} - \hat{S}_{n+1}) + (1 - \alpha_1)(\hat{\mu}_n + \hat{\beta}_n) \qquad (16)$$

où $\alpha_1 \in (0, 1)$ est une constante de lissage, associée à la mise-à-jour de la moyenne. Au temps n+1, nous avons aussi deux estimations possibles de la pente et nous utilisons une combinaison linéaire de celles-ci. D'une manière plus précise,

$$\hat{\beta}_{n+1} = \alpha_2(\hat{\mu}_{n+1} - \hat{\mu}_n) + (1 - \alpha_2)\hat{\beta}_n \qquad (17)$$

où $\alpha_2 \in (0, 1)$ est une constante de lissage, cette fois-ci associée à la mise-à-jour de la pente. Finalement, la mise-à-jour des ajustements saisonniers se fait par la récurrence

$$\hat{S}_{n+1} = \alpha_3(Z_{n+1} - \hat{\mu}_{n+1}) + (1 - \alpha_3)\hat{S}_{n-s+1} \qquad (18)$$

où $\alpha_3 \in (0, 1)$ est une constante de lissage, associée à la mise-à-jour des indices saisonniers. Les prévisions sont alors faites via

$$Z_n(js+k) = \hat{\mu}_n + \hat{\beta}_n(js+k) + \hat{S}_{n+k-s}, \qquad (19)$$

pour $j \geq 0$ et $k \in \{1, 2, \ldots, s\}$. Il faut cependant déterminer les valeurs initiales pour les récurrences impliquées. Le choix des constantes de lissage α_i, $1 \leq i \leq 3$, peut être fait en ayant comme objectif de minimiser l'erreur quadratique moyenne des prévisions d'horizon 1. Le lecteur voulant avoir plus de détails relativement au choix des valeurs initiales et des constantes α_i, $1 \leq i \leq 3$, est référé à la section 4.4.1 d'ABRAHAM et LEDOLTER (1983).

3.3.2 MODELE MULTIPLICATIF

Le modèle multiplicatif de Winters suppose que

$$Z_n = T_n \times S_n + \varepsilon_n \qquad (20)$$

où $T_{n+j} = \mu_{n+j} = \mu_n + \beta j$ est une tendance linéaire, tandis que S_n est une suite de constantes telles que $S_n = S_{n+s} = S_{n+2s} = \ldots$ pour $n \in \{1, \ldots, s\}$, qui sont en fait des ajustements, multiplicatifs cette fois, devant être faits aux valeurs T_n, afin de tenir compte du caractère cyclique de la série. On suppose de plus que $\sum_{i=1}^{s} S_i = s$. D'une manière analogue au cas additif, on trouve

$$\hat{\mu}_{n+1} = \alpha_1 \frac{Z_{n+1}}{\hat{S}_{n+1-s}} + (1 - \alpha_1)(\hat{\mu}_n + \hat{\beta}_n) \qquad (21)$$

$$\hat{\beta}_{n+1} = \alpha_2(\hat{\mu}_{n+1} - \hat{\mu}_n) + (1 - \alpha_2)\hat{\beta}_n \qquad (22)$$

$$\hat{S}_{n+1} = \alpha_3 \frac{Z_{n+1}}{\hat{\mu}_{n+1}} + (1 - \alpha_3)\hat{S}_{n-s+1} \qquad (23)$$

$$Z_n(js+k) = \{\hat{\mu}_n + \hat{\beta}_n(js+k)\}\hat{S}_{n+k-s}, \qquad (24)$$

pour $j \geq 0$ et $k \in \{1, 2, \ldots, s\}$. Le lecteur désirant plus de détails relativement au choix des valeurs initiales est référé à la section 4.4.2 d'ABRAHAM et LEDOLTER (1983).

Nous aimerions attirer l'attention du lecteur sur la simplicité des récurrences impliquées. Un programme, très compact et nécessitant peu de mémoire, peut être écrit en Pascal ou en Fortran, afin de déterminer les constantes α_i, $1 \leq i \leq 3$, les valeurs initiales et les prévisions. Une implémentation sur micro-ordinateur est tout aussi réalisable.

	Modèle additif sur Z_n	Modèle multiplicatif sur Z_n	Modèle additif sur $\ln Z_n$	Modèle multiplicatif sur $\ln Z_n$
α_1	0.100	0.300	0.166	0.300
α_2	0.010	0.100	0.010	0.010
α_3	0.300	0.100	0.300	0.010

Tableau 5.

Nous avons appliqué ces deux méthodes de lissage aux données brutes et au logarithme des données. Le tableau 5 présente les valeurs des constantes α_i, $1 \leq i \leq 3$, qui nous semblent être un bon choix.

3.4 Modèle proposé par Box et Jenkins

Au chapitre 9 de BOX et JENKINS (1976), il est suggéré un modèle de la forme

$$\nabla\nabla_{12} \ln X_t = (1 - \theta B)(1 - \Theta B^{12})a_t \qquad (25)$$

où $\{a_t\}$ est un bruit blanc. En utilisant les 132 premières observations, l'estimation des paramètres de ce modèle donne $\hat{\theta} = 0.34854$ et $\hat{\Theta} = 0.562207$.

4 COMPARAISON DES MODELES ET CONCLUSION

Sur une même série de 132 observations nous avons appliqué sept modèles différents. Comment choisir un modèle satisfaisant? Dans un premier temps, nous pouvons utiliser chacun des sept modèles afin de faire des prévisions pour la dernière année. Pour les modèles où nous avons utilisé le logarithme nous transformerons en sens inverse les données afin d'avoir une échelle commune pour les prévisions. On définit un indice, Q, afin d'évaluer la qualité des prévisions obtenues avec chacun des modèles.

$$Q = \frac{1}{12}\sum_{i=133}^{144}\left|\frac{Z_i - \hat{Z}_i}{Z_i}\right| \times 100 \qquad (26)$$

Il s'agit essentiellement de l'erreur relative moyenne. Le tableau 6 nous donne la valeur de Q pour chacun des modèles.

On constate que la régression avec variables indicatrices et la régression harmonique donnent les moins bons résultats. Bien que les R^2 soient effectivement élevés, il n'en demeure pas moins que les résidus de ces modèles ne semblent pas satisfaire les hypothèses de base, entre autres, ils sont fortement

autocorrélés. Le modèle additif de Winters appliqué au logarithme des données produit le meilleur indice Q. Il est suivi de près par le modèle classique de Box et Jenkins.

Il va de soi que cela ne démontre pas la supériorité de cette méthode sur les autres. Cependant, il s'agit d'un exemple où une méthode relativement simple donne d'aussi bons résultats que des méthodes beaucoup plus complexes.

Modèle	Q
Régression avec variables indicatrices	7.839
Régression harmonique	7.994
Modèle additif de Winters	4.667
Modèle multiplicatif de Winters	3.784
Modèle additif de Winters avec logarithme	2.732
Modèle multiplicatif de Winters avec logarithme	3.477
Modèle de Box et Jenkins	2.889

Tableau 6.

Lorsque se pose la question du choix d'une méthode de prévision, il faut peut-être prendre en compte la complexité de celle-ci afin d'évaluer les ressources nécessaires à son utilisation. Il est bien connu que les méthodes de lissage sont des méthodes simples et faciles à implémenter sur ordinateur.

Une étude préliminaire faite sur un échantillon d'une soixantaine de compagnies installées au Québec (Deslauriers, S., Fortin, A., et Latour, A., en préparation) et ayant pour but de comparer la qualité des prévisions faites à l'aide d'un modèle saisonnier de la classe ARIMA et de celles faites à partir d'un modèle multiplicatif de Winters indique que pour le type de séries considérées, aucune différence significative n'est décelée entre la performance de ces méthodes.

RÉFÉRENCES

ABRAHAM, B. et J. LEDOLTER (1983). Statistical Methods for Forecasting, Wiley, New York.

BOX, G. E. P. et G.M. JENKINS (1976). Times Series Analysis: Forecasting and control, 2ième édition, Holden-Day, San Francisco.

LA DETECTION DES OBSERVATIONS SURINFLUENTES DANS LES SERIES CHRONOLOGIQUES

(Version préliminaire)

Bernard Lefrançois
Division des mesures et de l'analyse des industries
Statistique Canada, Ottawa, Ontario K1A 0T6

1. RESUME

Supposons que $(X_t)_{t=1}^{n}$ soit une réalisation d'une série chronologique stationnaire (au sens faible), pour laquelle nous voulons obtenir un modèle ARMA. Pour ce faire, la fonction d'autocorrélation (FAC) échantillonnale $(r_k)_{k=1}^{L}$, où

$$r_k = c_k/c_0$$

et

$$c_k = \frac{1}{n} \sum_{t=1}^{n-k} (X_t-\bar{X})(X_{t+k}-\bar{X}),$$

joue un rôle essentiel dans l'identification, l'estimation et la validation du modèle ARMA. Il est alors très important d'identifier dès le début de la modélisation les observations qui auraient une trop grande influence sur le comportement des coefficients d'autocorrélation.

Dans cet article, nous décrivons comment obtenir des valeurs critiques pour déterminer si certaines observations d'une série chronologique stationnaire sont surinfluentes pour un coefficient d'autocorrélation donné, ou pour l'ensemble de la FAC échantillonnale. Cette méthode permet d'identifier avant la modélisation les observations qui pourraient faire problème.

2. CHOIX D'UNE MESURE D'INFLUENCE

Avant d'introduire les mesures d'influence qui seront utilisées dans le contexte des séries chronologiques, il est utile de se rapporter au développement de ces mesures dans le cas des échantillons aléatoires.

Tout d'abord, nous définissons la surinfluence comme suit: une observation est dite surinfluente pour une statistique T donnée, lorsque la contribution de cette observation à

l'évaluation de T est discordante par rapport à la contribution des autres observations du même échantillon.

Il est important de remarquer que par 'discordance', nous entendons ce qui est communément appelée 'aberrance'. Mais nous réservons ce dernier terme au cas bien précis d'une observation discordante par rapport au modèle statistique envisagé pour l'ensemble des observations.

Remarquons aussi qu'on ne peut définir l'influence d'une observation que par rapport à une statistique donnée.

Maintenant, supposons que la statistique qui nous intéresse soit calculée à partir d'un échantillon aléatoire $(X)_{i=1}^{n}$, c'est-à-dire que les X_i sont indépendantes et identiquement distribuées selon F, une loi quelconque. Supposons de plus que T puisse s'écrire comme une fonctionnelle de la loi F, c'est-à-dire qu'il existe une fonction ψ telle que $T[F] = \int \psi dF$. Hampel(1974) définit alors la <u>fonction d'influence</u> comme suit:

$$\mathcal{I}(x;T,F) = \lim_{h \to 0} m(x;T,F,h)$$

où $$m(x;T,F,h) = \frac{T[(1-h)F + h\delta_x] - T[F]}{h},$$

c'est-à-dire que la fonction d'influence de T est sa dérivée directionnelle selon la loi F.

En pratique, nous sommes intéressé à l'évaluation de cette fonction d'influence pour un échantillon observé. C'est ce qu'on appelle une <u>mesure d'influence</u>. Trois mesures d'influence sont les plus souvent utilisées. Ce sont:

La <u>mesure empirique d'influence additionelle</u>

$$EIC_i = \mathcal{I}(X_i;T,F_n)$$

où F_n est la fonction de répartition échantillonale (FRE).

La <u>mesure empirique d'influence réduite</u>

$$EIC(i) = \mathcal{I}(X_i;T,F_n^{(i)})$$

où $F_n^{(i)}$ est la FRE obtenue en supposant que la i^e observation est manquante.

La mesure d'influence échantillonnale

$$SIC_i = m(X_i;T,F_n,-(n-1)^{-1}) = (n-1)(T[F_n]-T[F_{n(i)}]).$$

Ces définitions sont les définition usuelles utilisées par Huber(1981) et Hampel et al.(1986). Or, il existe une relation importante entre ces trois mesures qui clarifie leur interprétation et qui nous permet d'en choisir une de préférence aux deux autres.

Cette relation fut démontrée par Pregibon(1981) et par Cook et Weisberg(1982) pour le cas très particulier de $\hat{\beta}$, l'estimateur des moindres carrés des paramètres du modèle linéaire usuel. Elle fut aussi démontrée par Critchley(1985) dans un contexte plus général. Nous l'avons démontrée sous des hypothèses moins restrictives que celles utilisées par Crithcley(1985), énoncées au théorème 1 ci-dessous.

Théorème 1. Soit $(Y_i)_{i=1}^n$ un échantillon de v.a. i.i.d., et soit T une statistique qui peut s'écrire comme fonctionnelle de la FRE F_n. Alors,

$$(n-1)EIC_i = \frac{\partial T}{\partial \omega_i}(1), \qquad nEIC(i) = \frac{\partial T}{\partial \omega_i}(0)$$

et

$$SIC_i = (n-1)\int_0^1 \frac{\partial T}{\partial \omega_i}\, d\omega_i$$

où ω_i est le poids donné à la i^e observation par F_n.

L'interprétation du théorème 1 est la suivante, et est identique à celle de Cook(1986). Chacune des mesures d'influence SIC_i, $EIC(i)$ et EIC_i constitue un sommaire de la variation dans la valeur de T induite par la variation du poids ω_i donné à la i^e observation par la FRE.

Ainsi, puisque la mesure d'influence échantillonnale SIC_i correspond à la moyenne de la courbe $\frac{\partial T}{\partial \omega_i}$, tandis que les autres mesures ne sont que des évaluations de cette même courbe à des points bien précis, il est préférable d'utiliser SIC_i comme mesure d'influence.

3. TRAVAUX PRECEDENTS

Dans le contexte des séries chronologiques, on ne peut démontrer le théorème 1 directement pour diverses raisons.

Une étant que l'hypothèse d'indépendance des observations est essentielle, puisque la définition usuelle de la fonction d'influence, indiquée à la section précédente, n'est valide que pour les échantillons aléatoires.

Pour les séries chronologiques, une autre définition doit être utilisée; voir Martin et Yohai(1986). Néanmoins, l'interprétation de la généralisation des fonctions d'influence par Martin et Yohai(1986) est identique à celle de la définition usuelle. Ce qui nous permet d'utiliser les résultats du théorème 1 comme définition des mesures d'influence pour les séries chronologiques.

Jusqu'à présent, nous n'avons retracé que quatre articles traitant de l'identification des observations surinfluentes pour la FAC: Chernick, Downing et Pike(1982), Lattin(1983), Li et Hui (1987) et Polasek(1982). Tous ces travaux ont en commun d'avoir adapté aux séries chronologiques des mesures d'influence pour les échantillons aléatoires, et de n'avoir donné aucun critère pour juger de la trop grande influence d'une observation.

Prenons par exemple l'article de Chernick, Downing et Pike(1982). Ils suggèrent d'utiliser $C^{+}_{i,k}$ pour mesurer l'influence de la i^e observation (X_i) sur le k^e coefficient d'autocorrélation (r_k), où:

$$C^{+}_{i,k} = \tilde{X}_i\tilde{X}_{i+k} - 0.5(\tilde{X}^2_i + \tilde{X}^2_{i+k})r_k,$$

avec

$$\tilde{X}_i = (X_i - \bar{X})/c_0^{1/2}.$$

Mais $C^{+}_{i,k}$, obtenue en utilisant une mesure de l'influence d'un couple d'observations sur le coefficient de corrélation linéaire usuel, ne tient pas compte de la contribution de X_{i-k} dans le calcul de r_k. De plus, ils suggèrent d'utiliser 1.0 comme valeur critique. Or, cette valeur critique ne correspond pas à la distribution asymptotique de $C^{+}_{i,k}$, et ils n'ont pas indiqué de quelle facon ils l'avaient obtenue.

Une fois la mesure d'influence choisie, l'obtention des critères, ou valeurs critiques, pour déterminer s'il y a surinfluence, correspond d'une manière formelle à un test d'hypothèse.

Quatre hypothèses nous intéressent tout particulièrement. La première consiste à déterminer si l'influence d'une observation <u>choisie à l'avance</u> sur un coefficient

d'autocorrélation donné est discordante, en ce sens que l'influence mesurée a une faible probabilité d'être observée selon la loi de probabilité de la mesure utilisée. La deuxième consiste à déterminer si l'influence d'une observation choisie à l'avance sur l'ensemble de la FAC est discordante.

La troisième consiste à déterminer si la plus grande influence observée pour un coefficient d'autocorrélation donné est beaucoup trop grande. Enfin, le dernier test consiste à déterminer si la plus grande influence observée pour l'ensemble de la FAC est beaucoup trop grande.

Pour effectuer ces tests d'hypothèse, il nous faut la distribution de $SIC_{i,k}$, ainsi qu'une mesure de l'influence d'une observation sur l'ensemble de la FAC et sa distribution.

4. DISTRIBUTION ASYMPOTIQUE DE $SIC_{i,k}$

On peut démontrer que

$$SIC_{i,k} = \frac{(\tilde{X}_i\tilde{X}_{i+k} + \tilde{X}_i\tilde{X}_{i-k} - r_k\tilde{X}_i^2)}{1 - \frac{\tilde{X}_i^2}{n-1}} + o_p(1),$$

c'est-à-dire que la mesure échantillonnale d'influence $SIC_{i,k}$ est constituée d'un quotient de v.a. dépendantes dont le numérateur est fonction du coefficient d'autocorrélation r_k, et d'un terme convergeant vers zéro en probabilité. Il serait donc malaisé d'obtenir la distribution exacte de $SIC_{i,k}$.

On doit alors utiliser le résultat asymptotique suivant:

Théorème 2. Soit $(X_t)_{t=1}^n$ une réalisation d'une série chronologique gaussienne stationnaire et régulière, centrée et réduite, ayant pour fonction d'autocorrélation $(\rho_k)_{k\in Z}$. Alors

$$SIC_{i,k} \xrightarrow{p} Sic_{i,k}$$

où la densité de $Sic_{i,k}$ est

$$f(s;\alpha,\sigma^2) = e^{\alpha s/\sigma^2}K_0(|s|/\sigma^2)/\pi\sigma,$$

où $K_\nu(x)$ est la fonction modifiée de Bessel de la 2e espèce d'ordre ν:

$$K_\nu(x) = \int_0^\infty e^{-x\cosh(t)}\cosh(\nu t)dt, \quad \text{pour } x > 0,$$

et où

$$\alpha = \begin{cases} \rho_k(\sigma^2+\rho_k^2)^{-1/2} & \text{pour } i=k+1(1)n-k \\ 0 & \text{autrement} \end{cases}$$

avec

$$\sigma^2 = \begin{cases} 2(1+\rho_{2k}-2\rho_k^2) & \text{pour } i=k+1(1)n-k \\ 1-\rho_k^2 & \text{autrement.} \end{cases}$$

Le théorème 2 nous permet d'obtenir les valeurs critiques asymptotiques pour déterminer si une observation choisie à l'avance est surinfluente pour un coefficient d'autocorrélation donné. Mais remarquons que la loi asymptotique de $SIC_{i,k}$ dépend de la FAC. Les valeurs critiques doivent donc être recalculées pour chaque série.

Dans la section suivante, nous transformerons le vecteur $SIC_{i,\cdot} = (SIC_{i,k})_{k=1}^{L}$ pour obtenir une mesure de l'influence d'une observation sur l'ensemble de la FAC. La densité asymptotique de cette mesure ne dépend pas de la FAC.

En transformant $SIC_{i,k}$ d'une façon analogue, les valeurs critiques asymptotiques obtenues peuvent alors être utilisées pour une quelconque série chronologique gaussienne stationnaire.

5. INFLUENCE SUR L'ENSEMBLE DE LA FAC

Pour mesurer l'influence de la i^e observation sur l'ensemble de la FAC, il faut réduire la longueur du vecteur $SIC_{i,\cdot}$ de sorte que le résultat mesure bien cette influence globale, et qu'il soit facile d'utilisation.

Tous les auteurs mentionnés précédemment ont suggéré d'utiliser comme mesure d'influence la somme des carrés des éléments indiqués par un '+' dans la matrice des influences montrée à la figure 1. C'est ce que nous appelons le crochet d'influence.

Figure 1. Le crochet d'influence

	Décalages							
Obs	1	2	3	4	5	6	...	L
.				.				
.				.				
.				.				
i-4	.	.	.	+	.	.	...	.
i-3	.	.	+	.	.	.	...	.
i-2	.	+	.	.	.	.	...	.
i-1	+	.	.	.	.	.	...	.
i	+	+	+	+	+	+	...	+
i+1	.	.	.	.	.	.	...	.
i+2	.	.	.	.	.	.	...	.
.				.				
.				.				
.				.				

Or, ceci occasionne deux difficultés. La première est que la partie supérieure du crochet (la diagonale dans la figure 1) ne correspond pas à l'influence de la i^e observation sur un quelconque coefficient d'autocorrélation. La deuxième difficulté est que la distribution asymptotique de cette mesure est inutilement compliquée.

En fait, nous suggérons d'utiliser la somme des carrés des éléments <u>centrés et réduits</u> du vecteur $SIC_{i,.}$, ce qui a comme avantage de ne mesurer que l'influence de la i^e observation, et, comme le montre le théorème 3 ci-dessous, la statistique résultante Q_i a une distribution asymptotique relativement simple.

Avant d'énoncer le théorème 3, il faut définir certaines quantités. Pour L et n des constantes, posons

$$h(i) = min\{i, n-i, L\}$$

et $$J_{h(i)} = diag(1'_{h(i)},\ 0_{1x(L-h(i))}),$$

où 1_m représente le vecteur-colonne de longueur m dont chaque élément est l'unité.

Posons $$R'_{k,l} = (r_k,\ r_{k+1},\ \ldots\ ,\ r_l),$$

$$T_{k,l} = \begin{pmatrix} r_k & r_{k+1} & r_{k+2} & \cdots & r_l \\ r_{k+1} & r_k & r_{k+1} & \cdots & r_{l-1} \\ r_{k+2} & r_{k+1} & r_k & \cdots & r_{l-2} \\ \vdots & & \ddots & & \vdots \\ r_l & r_{l-1} & r_{l-2} & \cdots & r_k \end{pmatrix},$$

et

$$H_{k,l} = \begin{pmatrix} r_k & r_{k+1} & r_{k+2} & \cdots & r_{\frac{l}{2}} \\ r_{k+1} & r_{k+2} & r_{k+3} & \cdots & r_{\frac{l}{2}+1} \\ r_{k+2} & r_{k+3} & r_{k+4} & \cdots & r_{\frac{l}{2}+2} \\ \vdots & & \ddots & & \vdots \\ r_{\frac{l}{2}} & r_{\frac{l}{2}+1} & r_{\frac{l}{2}+2} & \cdots & r_l \end{pmatrix}.$$

$T_{k,l}$ est une matrice de Toeplitz tandis que $H_{k,l}$ est une matrice de Hankel. Toutes deux sont symétriques. Finalement, posons

$$\tilde{T}_{k,l} = T_{k,l} - r_k R_{1,l-k+1} R'_{1,l-k+1}$$

et

$$\tilde{H}_{k,l} = H_{k,l} - r_{k-2} R_{1,\frac{l}{2}-k+1} R'_{1,\frac{l}{2}-k+1}.$$

Nous pouvons alors démontrer le théorème suivant.

<u>Théorème 3</u>: Soit $(X_t)_{t=1}^{n}$ une série chronologique telle que décrite au théorème 2. Pour L constant par rapport à n, et pour $n \geq 2L+1$, i=1(1)n, définissons

$$\mu_i = J_{h(i)} R_{1,L} X_i$$

et

$$\Sigma_i = \tilde{T}_{0,L-1} + J_{h(i)} \tilde{T}_{0,L-1} J_{h(i)} + \tilde{H}_{2,2L} J_{h(i)} + J_{h(i)} \tilde{H}_{2,2L}.$$

Alors la densité asymptotique de

$$Q_i = (SIC_{i,.} - X_i\mu_i)'\Sigma_i^{-1}(SIC_{i,.} - X_i\mu_i)$$

est

$$f(q;L) = \frac{q^{\frac{L-3}{4}} K_{\frac{L-1}{2}}(q^{1/2})}{2^{\frac{L-1}{2}} \pi^{1/2} \Gamma(\frac{1}{2})}$$

où $K_\nu(x)$ est la fonction modifiée de Bessel de la deuxième espèce d'ordre ν.

La densité asymptotique de Q_i ne dépend donc que du nombre de coefficients d'autocorrélation sur lesquels l'influence est mesurée simultanément.

6. TEST DU MAXIMUM D'INFLUENCE

En pratique, nous serons plus souvent intéressé à vérifier si la plus grande influence observée est trop grande, que de tester l'influence d'une observation choisie à l'avance. Nous ne traiterons ici que de l'influence des observations sur l'ensemble de la FAC.

Pour obtenir les valeurs critiques exactes, nous devrions obtenir la distribution de

$$Q_{(n)} = \max_{i=1(1)n} (Q_i).$$

Puisque les Q_i sont dépendants, il serait difficile d'obtenir la distribution exacte de $Q_{(n)}$. Même pour le processus asymptotique $(Q_i)_{i=1}^n$, défini pour chaque i=1(1)n

par

$$Q_i = \operatorname{plim}_{n \to \infty} Q_i,$$

la loi de $Q_{(n)}$ est aussi fort complexe.

Nous devons donc utiliser des inégalités du type de Bonferroni pour obtenir des bornes sur les valeurs critiques de $Q_{(n)}$. Les bornes que nous avons obtenues sont celles pour $Q_{(n)}$. Lorsque la série chronologique étudiée est observée suffisamment longtemps, les bornes pour les valeurs critiques asymptotiques devraient aussi être applicables aux valeurs critiques exactes.

Etant donné les événements $E_1, E_2, \ldots, E_n$, et $E = \bigcup_{i=1}^{n} E_i$, les <u>première et deuxième inégalités de Bonferroni</u> sont:

$$\mathcal{B}_2 = \sum_{i=1}^{n} P(E_i) - \sum_{i<j} P(E_i \cap E_j) \leq P(E) \leq \sum_{i=1}^{n} P(E_i) = \mathcal{B}_1.$$

Hunter(1976) obtint une borne supérieure plus précise, décrite au théorème 4.

<u>Théorème 4. Inégalité de Hunter</u>:

Considérons les événements $(E_i)_{i=1}^{n}$ comme les sommets d'un graphe G, reliés par les arcs g_{ij} si et seulement si $P(E_i \cap E_j) > 0$. Soit T un sous-graphe de G. Alors,

$$P(E) = P(\bigcup_{i=1}^{n} E_i) \leq \sum_{i=1}^{n} P(E_i) - \sum_{((i,j):g_{ij} \in T)} P(E_i \cap E_j) = \mathcal{B}_3$$

si et seulement si T est un arbre. De plus, l'égalité a lieu lorsque G est lui-même un arbre et T = G.

Il est clair que le supremum de P(E) est obtenu lorsque T est la plus grande arborescence de G, que l'on trouve par l'application de l'algorithme de Kruskal(1956).

Utilisant l'inégalité $\mathcal{B}_3$ de Hunter et l'inégalité $\mathcal{B}_2$ de Bonferroni, et en supposant que le processus $(Q_i)_{i=1}^{n}$ est strictement stationnaire, on démontre que

$$I_{n-1}(c) \leq P(Q_{(n)} > c) \leq S_{n-1}(c)$$

où
$$I_{n-1}(c) = n\left[p_0(c) - \sum_{k=1}^{n-1} \left(1 - \frac{k}{n}\right) p_k(c) \right],$$

$$S_{n-1}(c) = n\left[p_0(c) - \sum_{k=1}^{m} \left(1 - \frac{\sigma(k)}{n}\right) p_{(n-k)}(c) \right]$$

avec $p_{(k)}(c) = p_{\sigma(k)}(c)$ et $p_{(n-1)}(c) \geq \ldots \geq p_{(2)}(c) \geq p_{(1)}(c)$,

et m tel que
$$1 = \frac{1}{n} + \sum_{i=1}^{m} \left(1 - \frac{\sigma(n-i)}{n}\right)$$

et où $p_k(c) = P(Q_i > c,\ Q_{i+k} > c)$, $p_0(c) = P(Q_i > c)$.

On obtient donc un test qui s'utilise de la même façon que celui de Durbin et Watson(1950).

Trois difficultés cependant doivent être résolues pour obtenir en pratique les valeurs critiques.

La première est que le processus $(Q_i)_{i=1}^{n}$ n'est pas strictement stationnaire. Mais la portion centrale du même processus, $(Q_i)_{i=L+1}^{n-L}$, est stationnaire. De plus, la contribution des termes non stationnaires dans l'évaluation des bornes $I_{n-1}(c)$ et $S_{n-1}(c)$ est constante par rapport à la longueur de la série. Donc pour les séries observées suffisamment longtemps, l'hypothèse de la stationnarité de l'ensemble de la série $(Q_i)_{i=1}^{n}$ n'est pas trop critique.

Deuxièmement, on ne peut calculer les probabilités conjointes $p_k(c)$ jusqu'à k=n-1. Pour en obtenir une estimation, on suppose que la série observée $(X_t)_{t=1}^{n}$ est un processus de moyenne mobile d'ordre 2L, où L est le nombre de coefficients de la FAC calculée. Avec cette hypothèse, $p_k(c) = p_0(c)^2$ aussitôt que $k > 4L$. Cette hypothèse n'est pas trop restrictive, puisque l'identification d'un processus ARMA à partir des L premières autocorrélations repose implicitement sur l'hypothèse que la série observée est un processus de moyenne-mobile d'ordre au plus L.

Finalement, la plus importante difficulté est qu'il n'existe pas encore d'expression numériquement utile pour les probabilités conjointes $p_k(c)$, pour k=1(1)4L. Nous suggérons alors d'utiliser une approximation log-normale.

7. CONCLUSION

La méthode que nous avons exposée ici semble fonctionner à souhait sur les quelques séries que nous avons déjà analysées. Certaines questions, cependant, demeurent sans réponse pour le moment. Par exemple, pour obtenir une bonne approximation de la probabilité conjointe $p_k(c)$, il faudrait obtenir une bonne approximation de la probabilité conjointe de deux formes quadratiques non-centrées et dépendantes.

8. BIBLIOGRAPHIE

Chernick, Downing et Pike(1982) *J. Amer. Statist. Assoc.*, **77**, 380, pp. 743-747

Cook(1986) *J. Roy. Statist. Soc. Ser. B*, **48**, 2, pp. 133-169

Cook et Weisberg(1982) *Residuals and Influence in Regression*, Chapman and Hall, Londre

Critchley(1985) *Biometrika*, **72**, 3, pp. 627-636

Durbin et Watson(1950) *Biometrika*, **37**, pp. 409-428

Hampel(1974) *J. Amer. Statist. Assoc.*, **69**, pp. 383-393

Hampel, Ronchetti, Rousseeuw et Stahel(1986) *Robust Statistics. The Approach Based on Influence Functions*, J. Wiley, New York

Huber(1981) *Robust Statistics*, J. Wiley, New York

Hunter(1976) *J. Appl. Prob.*, **13**, pp. 597-603

Kruskal(1956) *Proc. Amer. Math. Soc.*, **7**, pp. 48-50

Lattin(1983) in *Computer Science and Statistics: Proceedings of the 14th Symposium on the Interface*, Heiner, K.W. et al. (Eds.), Springer-Verlag, Berlin, pp. 283-286

Li et Hui(1987) in *Amer. Statist. Assoc. Proc. Bus. Econom. Statist. Section*, pp. 465-468

Martin et Yohai(1986) *Ann. Statist.*, **14**, pp. 781-855

Polasek(1982) *Robust Estimation and Resistance Analysis for the Auto-Correlation Function*, Preprint #47, Institute of Statistics and Computer Science, Université de Vienne, Autriche.

Pregibon(1981) *Ann. Statist.*, **9**, 4, pp. 705-724

L'AJUSTEMENT DES DONNÉES ANNUELLES DE FLUX À L'ANNÉE CIVILE

Claude Poirier et Normand Laniel
Statistique Canada, 11-K immeuble R.H. Coats, Ottawa (Ont) K1A 0T6

L'utilisation de données basées sur les années financières des entreprises plutôt que sur l'année civile pose un problème d'inconsistance des estimations annuelles. Deux types d'ajustements sont disponibles afin de réduire cette inconsistance: les ajustements qui utilisent des données sous-annuelles et ceux qui n'en utilisent pas. Une étude visant à évaluer la performance de chacune de ces méthodes a été menée à Statistique Canada. Un compte-rendu de cette étude est présenté dans ce document.

MOTS CLÉS : Correction de données, période financière, enquête annuelle

1. INTRODUCTION

Le présent rapport vise à identifier le problème de l'ajustement à l'année civile dans le contexte du projet de remaniement des enquêtes-entreprises à Statistique Canada. Il vise aussi à proposer des procédures permettant de corriger l'incompatibilité avec l'année civile, des statistiques annuelles basées sur l'année financière des entreprises. Ces procédures devraient être simples et efficaces et devraient pouvoir s'appliquer à toutes les enquêtes annuelles de production économique.

Ce document ne traite que quelques ajustements applicables à des données de flux (valeurs accumulées sur un intervalle de temps). Laniel et Poirier (1989) donnent plus de détails sur ces ajustements et sur ceux applicables à des données de stock (valeurs mesurées en un point spécifique).

2. DÉFINITION DU PROBLÈME

Dans le cadre du projet de remaniement des enquêtes-entreprises (PREE) à Statistique Canada, c'est une politique pour toutes les enquêtes annuelles que d'utiliser l'année financière du répondant comme période de déclaration. Ceci vient du fait que les données économiques d'intérêt sont plus facilement disponibles pour l'année financière que pour tout autre période, et que les données ainsi obtenues sont plus exactes car les entreprises font souvent des ajustements à leurs livres comptables à la fin des exercices financiers. De plus, cette politique permet de réduire le fardeau du répondant et l'erreur de réponse.

Sous l'égide du PREE, il a été décidé que les données d'enquêtes pour l'année de référence Y seraient obtenues auprès des entreprises ayant une année financière se terminant entre le 1er avril de l'année civile Y et le 31 mars de l'année civile Y+1 (Roberge 1988). Ce choix est un compromis entre le désir d'un court délai de production pour les statistiques annuelles et la maximisation de la couverture de l'année civile Y.

Cependant, cette politique entraine que les données annuelles ne correspondent pas nécéssairement à l'année civile visée. Ces données peuvent correspondre à des périodes décalées par rapport à l'année civile ou même d'une durée moindre que douze mois. Ces deux facteurs peuvent biaiser les estimations. Par exemple, des années financières prennant fin avant le 31 décembre de l'année visée contribuent à sous-estimer les ventes en période de croissance économique ou à les sur-estimer en période de récession économique. Des années financières de moins de douze mois contribuent aussi à sous-estimer les ventes. Des biais de 5% et même de 10% pour certains groupes d'industries ont été observés sur les données fournies par les manufacturiers canadiens. Il est donc souhaitable d'ajuster les statistiques annuelles basées sur les périodes financières des entreprises parce que peu importe les avertissements accompagnant les données diffusées, les utilisateurs seront portés à croire que les statistiques annuelles font référence à l'année civile.

3. LES MÉTHODES

Cette section vise à dresser un inventaire de méthodes visant à ajuster des données annuelles de flux basées sur l'année financière des entreprises pour qu'elles correspondent à l'année civile. Ces méthodes peuvent se diviser en deux groupes. Un premier groupe est formé des méthodes n'utilisant que des données annuelles et un deuxième groupe est formé de celles qui utilisent à la fois des données annuelles et mensuelles.

3.1 Méthodes n'utilisant que des données annuelles

Les méthodes n'utilisant que des données annuelles se différentient entre elles par le nombre d'années pour lesquelles des données annuelles financières sont nécéssaires à l'ajustement et aussi par le niveau de totalisation de ces données.

A) Statu quo

Une première méthode est celle utilisée actuellement à Statistique Canada et consiste à ne faire aucune transformation des données obtenues des entreprises. Cette méthode peut être qualifiée de méthode Identité ou de Statu quo. Si on représente par:

X(Y): Le total (ou une estimation) des données financières* pour l'année de référence Y,

on obtient un ajustement $X^a(Y)$ pour l'année Y comme suit:

$$X^a(Y) = X(Y)$$

Il est raisonnable de penser que cette méthode donnerait de bons résultats lorsqu'on est en présence d'une économie stagnante (c.-à-d. sans croissance) où toutes les entreprises ont une année financière d'une durée de douze mois. Cette méthode servira de référence lors de l'évaluation des autres méthodes.

---

* L'expression "données financières" est utilisé dans ce document pour signifier "données basées sur l'année financière des entreprises".

B) Moyenne de macro-données annuelles

Comme le faisait le ministère du revenu national (1964), il s'agit ici de considérer des totalisations de données financières pour deux années consécutives et d'en faire un ajustement basé sur un principe de linéarité de la croissance économique. Les totalisations, qu'on nomme aussi macro-données, sont définies ici sur des domaines correspondant au mois de fin de la période financière. Ainsi en posant:

$X_k(Y)$: le sous-total (ou une estimation) des données annuelles financières pour les unités ayant une période financière se terminant le k^e mois de l'année civile Y, si $4 \leq k \leq 12$, et le k^e mois de l'année civile Y+1, si $1 \leq k \leq 3$, (soit les données pour l'année de référence Y),

on obtient un ajustement $X_k^a(Y)$ à partir d'une moyenne de deux macro-données consécutives $X_k(Y-1)$ et $X_k(Y)$ comme suit:

$$X_k^a(Y) = \alpha_k X_k(Y-1) + (1-\alpha_k) X_k(Y)$$

où $$\alpha_k = \begin{cases} k/12 & \text{si } 1 \leq k \leq 3 \\ (k-12)/12 & \text{si } 4 \leq k \leq 12 \end{cases} .$$

Cet ajustement est une interpolation linéaire lorsque $1 \leq k \leq 3$ et une extrapolation linéaire lorsque $4 \leq k \leq 12$. Remarquons que le même principe peut s'appliquer lorsqu'on utilise les macro-données de Y et Y+1 pour faire l'ajustement. Cette variante est utile lorsque les données de Y-1 ne sont pas disponibles. Cependant, le calcul des valeurs ajustées s'en trouverait retardé puisqu'il faudrait attendre les données de Y+1.

Une fois les ajustements effectués pour chacun des sous-totaux $X_k(Y)$, on obtient la valeur ajustée du total X(Y) par la somme suivante:

$$X^a(Y) = \sum_{k=1}^{12} X_k^a(Y) .$$

Ces moyennes de macro-données supposent que les entreprises ont toutes des années financières de douze mois et qu'elles sont en opération pendant chacun des douzes mois de l'année civile Y. Malheureusement, ces deux hypothèses sont souvent rejetées en pratique, notamment dans le cas de naissance et de mort d'entreprise. C'est pourquoi il faut aussi considérer des méthodes s'appliquant aux micro-données, permettant ainsi de satisfaire des hypothèses plus réalistes. De telles méthodes sont décrites ci-après.

C) Moyenne de micro-données annuelles

Il s'agit ici d'effectuer les ajustements directement au niveau du répondant ou micro-données. Ceci permet de prendre en considération la durée de l'année financière, son chevauchement avec l'année civile d'intérêt et le nombre de mois où l'entreprise fut en opération durant cette dernière. Comme pour la moyenne de macro-données, ce type d'ajustement est basé sur l'hypothèse d'une croissance économique constante. En posant:

$x_i(Y)$: la micro-donnée financière de l'unité i, pour l'année de réf. Y,

l'ajustement par moyenne des micro-données Y-1 et Y peut se représenter sous la forme:

$$\overset{a}{x}_i(Y) = \alpha_i\, x_i(Y\text{-}1) + \beta_i\, x_i(Y)$$

où α_i et β_i sont fonction de la fin des années financières Y-1 et Y, ainsi que de leur durée. Ces paramètres sont déterminés en identifiant premièrement une suite de valeurs mensuelles $z_i(M) = a_i + b_i \cdot M$ à croissance linéaire telle que leurs sommes sur les mois M de l'année financière Y-1 et de l'année financière Y soient respectivement égales aux données financières $x_i(Y\text{-}1)$ et $x_i(Y)$. Ainsi, α_i et β_i sont tels que l'ajustement est égale à la somme des $z_i(M)$ sur les mois de l'année civile au cours desquels l'entreprise était en activité. Les M utilisés ici ne sont qu'une représentation chronologique des mois d'intérêt sur l'ensemble des nombres entiers.

Tout comme pour la moyenne de macro-données, l'ajustement peut être utilisé avec les données de Y et Y+1 lorsque celle de Y-1 n'est pas disponible. Cependant, encore une fois, un délai dans le calcul de l'ajustements serait introduit.

Une fois les ajustements effectués pour chacun des répondants, on obtient la valeur ajustée du total X(Y) par une des sommes suivantes:

$$\overset{a}{X}(Y) = \begin{cases} \sum_{i\in\Omega} \overset{a}{x}_i(Y) & \text{si l'enquête annuelle est un recensement} \\ \sum_{i\in S} w_i \overset{a}{x}_i(Y) & \text{si l'enquête annuelle est un sondage} \end{cases}$$

Toutefois un tel ajustement suppose une saisonnalité nulle. Cette hypothèse n'était pas nécessaire au niveau des macro-données puisque la saisonnalité n'a aucun effet lorsqu'on utilise toujours des données se rapportant à douze mois consécutifs.

D) <u>Ajustement pour la durée</u>

L'ajustement pour la durée est une variante de la moyenne de micro-données annuelles. Cette méthode d'ajustement ne considère que les données de l'année de référence Y. Elle peut donc être utile lorsque les données de Y-1 et Y+1 ne sont pas disponibles. L'ajustement consiste à appliquer un facteur de correction tenant compte de la durée de l'année financière et du nombre de mois d'opération durant l'année civile. Il fait donc l'hypothèse d'une économie stagnante et sans saisonnalité. Ainsi, la valeur ajustée pour une micro-donnée est donnée par:

$$\overset{a}{x}_i(Y) = \beta_i\, x_i(Y)$$

où $$\beta_i = \frac{\text{nb de mois en activité durant l'année civile Y}}{\text{durée de la période financière Y}}$$

Exemple: Une unité qui meurt le 31 août de l'année Y et qui se retrouve avec une période financière de 11 mois, se verra ajuster sa valeur annuelle $x_i(Y)$ par:

$$\overset{a}{x}_i(Y) = 8/11\ x_i(Y)\ .$$

3.2 Méthodes utilisant des données annuelles et mensuelles

Il y a plus d'une façon de combiner les données mensuelles avec les données annuelles pour en faire des ajustements à l'année civile. Entre autre variante les données mensuelles peuvent être utilisées au niveau micro ou macro tout comme les données annuelles. Aussi les données mensuelles peuvent provenir d'une enquête mensuelle correspondante à l'annuelle ou encore d'une autre source. Deux variantes sont décrites et discutées ici.

A) Ajustement de macro-données annuelles avec macro-données mensuelles

Une première façon d'utiliser des données mensuelles pour ajuster les données annuelles est de considérer des totalisations de données annuelles financières ainsi que des totalisations de données mensuelles. Ces macro-données sont encore une fois liées au mois de fin de la période financière.

Soit $Z_k(M)$: le sous-total (ou une estimation) des données mensuelles du mois M pour les unités ayant une période financière se terminant le k^e mois de l'année civile Y, si $4 \leq k \leq 12$, et le k^e mois de l'année civile Y+1, si $1 \leq k \leq 3$,

on obtient un ajustement pour $X_k(Y)$ en utilisant un quotient comme suit:

$$\overset{a}{X}_k(Y) = R_k(Y) \cdot X_k(Y)$$

où

$$R_k(Y) = \frac{\sum_{M \in \text{année civile Y}} Z_k(M)}{\sum_{M \in \text{année finan. Y se terminant au mois k}} Z_k(M)}$$

et finalement la somme des $\overset{a}{X}_k(Y)$ fournit une valeur ajustée pour X(Y).

Lorsque les macro-données mensuelles proviennent d'une enquête par échantillon, on doit s'attendre à ce que cette méthode engendre une croissance de la variance. Aussi, la différence existant entre les données mensuelles et la valeur annuelle financière correspondante pour une même entreprise, peut contribuer à biaiser les ajustements. Cette différence est entre autres due à des ajustements comptables effectués par les entreprises sur leurs données annuelles ou encore simplement à une mauvaise qualité des données mensuelles fournies par les entreprises. Une autre source de biais vient du fait que l'utilisation de macro-données suppose que les années financières ont toutes une durée de douze mois et que les entreprises sont en opération pendant chacun des douzes mois de l'année civile.

B) Ajustement de micro-données annuelles avec micro-données mensuelles

Il s'agit ici d'utiliser des données mensuelles pour ajuster les données annuelles à un niveau micro. Cette utilisation de micro-données mensuelles se fait comme suit:

$$\overset{a}{x}_i(Y) = r_i(Y) \cdot x_i(Y)$$

où $$r_i(Y) = \frac{\sum_{M \in \text{l'année civile Y}} z_i(M)}{\sum_{M \in \text{année financière Y de l'unité i}} z_i(M)}$$

Un problème se pose lorsque, pour une unité, les données mensuelles $z_i(M)$ ne sont pas disponibles pour certains mois dû à des rotations échantillonnales mensuelles. Dans ce cas, une imputation par la tendance sera effectuée à l'aide de la macro-donnée $Z_k(M)$ du groupe k ayant une même fin d'exercice financier que l'unité en question.

4. ÉVALUATION EMPIRIQUE

4.1 Buts

Le but principal de l'évaluation empirique est de trouver la ou les méthodes d'ajustement permettant le mieux de réduire le biais du statu quo tout en n'augmentant pas la variance échantillonnale des données annuelles. Cette évaluation doit se faire dans un contexte identique à celui où les méthodes seraient utilisées si elles étaient implantées. C'est pourquoi les méthodes ont été évaluées au moyen d'une simulation de type Monte-Carlo, ceci à partir d'une centaine d'échantillons mensuels et annuels tirés de données réelles. Les données annuelles provenaient du recensement annuel des manufactures tandis que les données mensuelles provenaient de l'enquête mensuelle sur les livraisons, stocks et commandes. Des données de flux, représentant des livraisons, étaient retenues pour 21 sous-populations définies par code d'activité économique et par province. Chacune des méthodes d'ajustement a ensuite été évaluée pour chacune des sous-populations pour l'année de référence 1984.

Pour ce qui est des plans d'échantillonnage et des estimateurs utilisés, ils correspondent à la pratique des enquêtes économiques impliquées dans le projet de remaniement des enquêtes-entreprises et ce autant annuellement que mensuellement.

Ainsi, sur une base annuelle, un plan d'échantillonnage stratifié de Bernouilli à deux phases a été utilisé avec un estimateur par double quotient utilisant la variable auxiliaire unité. Sur une base mensuelle, un échantillonnage stratifié avec rotations mensuelles par groupes a été utilisé pour faire en sorte qu'aucune unité ne soit échantillonnée plus de 24 mois consécutifs ou mise hors de l'échantillon pour une période inférieure à 12 mois. L'estimateur associé était celui de Mickey et utilisait la variable auxiliaire représentant la taille des groupes. Des détails sur ces plans d'échantillonnages et sur ces stratégies d'estimation sont données par Choudhry, Lavallée et Hidiroglou (1989), Dumais et Carpenter (1989) et Hidiroglou et Choudhry (1989).

4.2 Comparaison des méthodes

Il est évident qu'une méthode d'ajustement à l'année civile doive réduire le biais autant que possible puisque c'est là sa raison d'être. Cependant, cela doit être fait sans qu'une autre source d'erreur d'ampleur égale au biais à corriger ne soit introduite car, autrement, l'exercice serait inutile. L'évaluation des méthodes devait donc considérer le rendement de chaque méthode au point de vue de l'écart

quadratique moyen ou de sa racine carrée. Cet écart quadratique moyen (EQM) pouvait être évalué grâce à des données théoriques fabriquées à partir d'un processus complexe visant à recréer les phénomènes observés en réalité. Les détails de cette méthodologie sont donnés par Laniel et Poirier (1989).

Pour les raisons décrites au paragraphe précédent, les méthodes d'ajustement ont toutes été comparées au statu quo en utilisant la différence des $\sqrt{EQM}$ ainsi que la variance de cette différence. Notons que cette variance a pu être estimée par jackknife à partir des 100 réalisations d'ajustements obtenues des répétitions d'échantillons. Les résultats de cette évaluation sont donnés en appendice.

4.3 Conclusions

Il ressort des résultats obtenus en appendice et d'une analyse plus détaillée des biais, variances et EQM des méthodes, qu'il n'est pas avantageux d'effectuer des ajustements à un niveau macro ou d'utiliser des données mensuelles.

La principale cause du biais introduit par le statu quo vient de la différence entre la longueur des années financières et le nombre de mois en activité durant l'année civile, et non pas du décalage des années financières par rapport à l'année civile. Ceci explique le mauvais rendement des méthodes agissant à un niveau macro, qui n'ajustent que pour le décalage. Aussi, les méthodes utilisant des données mensuelles ont montré un rendement décevant. Ceci est dû soit à une augmentation de la variance provenant des estimations mensuelles, soit à une augmentation du biais provenant des imputations effectuées pour des micro-données manquantes, soit à la mauvaise qualité des séries mensuelles, soit encore à une combinaison de ces trois facteurs.

Ainsi, pour ajuster des données de flux, il est recommandé de n'utiliser qu'un ajustement pour la durée tel que décrit à la section 3.1, même si des données annuelles sont disponibles pour les années précédente et suivante et même si des données mensuelles sont disponibles.

APPENDICE

RÉSULTATS DE L'ÉVALUATION EMPIRIQUE POUR DES DONNÉES DE FLUX

Il est évident qu'une méthode d'ajustement devrait, en moyenne, résulter en un EQM moindre que celui du statu quo. Mais pour garder une certaine fiabilité des estimations il a été décidé que, pour être considérée efficace, une méthode devait en tout temps entrainer un EQM moindre ou égale à celui du statu quo. Ainsi en posant l'hypothèse nulle suivante:

$$H_o: \sqrt{EQM_{méthode}} - \sqrt{EQM_{statuquo}} \leq 0 ,$$

l'EQM d'une méthode est considéré supérieur à celui du statu quo si:

i) H_o est rejetée avec un test t, et

ii) $$\frac{\sqrt{EQM_{méthode}} - \sqrt{EQM_{statuquo}}}{\text{Total théorique}} > c$$

où c est une constante prédéterminée. Ces deux conditions permettent d'identifier les différences significatives i) au point de vue statistique et ii) au point de vue pratique. Notons que le choix d'une valeur précise pour la constante c est difficile, c'est pourquoi, pour les besoins de la simulation, différentes valeurs y ont été assignées. Elles sont .000, .005, .010, .015, .020 et .025.

À partir des deux conditions énoncées plus haut, on peut comptabiliser, pour chacune des méthodes, le nombre de populations parmi les 21 où l'EQM est considéré supérieur à celui du statu quo. On obtient:

Nombre de populations, parmi les 21, où les ajustements entrainent un EQM <u>supérieur</u> à celui du STATU QUO

Méthode	c:	.000	.005	.010	.015	.020	.025
Moyenne de macro-données annuelles		12	10	7	7	5	5
Moyenne de micro-données annuelles		10	9	7	5	3	3
Ajustement pour la durée (micro-données)		4	0	0	0	0	0
Ajust. de macro-données avec données mensuelles		12	10	8	7	7	7
Ajust. de micro-données avec données mensuelles		11	8	6	5	4	3

Ces résultats montrent que l'ajustement le plus fiable est celui pour la durée. En effet, lorsqu'on est prêt à accepter une différence des $\sqrt{EQM}$ de 0.5% (c=0.005), la méthode n'est jamais pire que le statu quo.

Maintenant, avant de consacrer de l'énergie dans l'implantation d'un système d'ajustement comme celui pour la durée, il reste à s'assurer que, même s'il n'est jamais pire que le statu quo, il soit meilleur dans au moins quelques cas. C'est ainsi qu'en inversant les signes d'inégalité dans les règles de décision au début du présent appendice, on peut vérifier, pour chacune des méthodes, si elle entraine un EQM moindre que celui du statu quo. Avec les résultats qui suivent, on remarque que l'ajustement pour la durée rencontre les conditions exigées.

Nombre de populations, parmi les 21, où les ajustements entrainent un EQM inférieur à celui du STATU QUO

Méthode	c: -.000	-.005	-.010	-.015	-.020	-.025
Moyenne de macro-données annuelles	8	7	4	4	4	3
Moyenne de micro-données annuelles	10	7	5	5	5	4
Ajustement pour la durée (micro-données)	9	3	3	3	3	3
Ajust. de macro-données avec données mensuelles	5	4	4	2	2	2
Ajust. de micro-données avec données mensuelles	7	6	5	5	5	5

BIBLIOGRAPHIE

[1] Choudhry, G.H., Lavallée, P. et Hidiroglou, M.A. (1989), "Two-Phase Sample Design for Tax Data", Projet de remaniement des enquêtes-entreprises, Statistique Canada.

[2] Dumais, J. et Carpenter, R. (1989), " Methodology of the Generalized Estimation System for Sub-Annual Business Surveys", Statistique Canada.

[3] Hidiroglou, M.A. et Choudhry, G.H. (1989), "Generalized Sampling and Estimation Methodology for Sub-Annual Business Surveys at Statistique Canada", Division des méthodes d'enquêtes-entreprises, Statistique Canada.

[4] Laniel, N. et Poirier, C. (1989), "Ajustement des données d'enquêtes annuelles à l'année civile", Projet de remaniement des enquêtes-entreprises, Statistique Canada.

[5] Ministère du revenu national, division de l'impôt (aujourd'hui Revenu Canada) (1964), "Statistique fiscale 1964, partie 2", Ottawa.

[6] Roberge, D. (1989), "Target Population, Definition, Identification and Extraction", Projet de remaniement des enquêtes-entreprises, Statistique Canada.

PROBLÈMES PRATIQUES ASSOCIÉS AU CALCUL DE LA VARIANCE DES DONNÉES DÉSAISONNALISÉES PAR X-11-ARIMA

Benoit Quenneville et Estela Bee Dagum
Division de la recherche et de l'analyse des chroniques
Statistique Canada

1. INTRODUCTION.

La demande pour les estimés d'écart-type des données désaisonnalisées date de 1962 alors qu'un comité Présidentiel des États-Unis recommande *"that estimates of the standard errors of seasonally adjusted data be prepared and published as soon as the technical problems have been solved"*. Dix sept ans plus tard, une commission nationale reprend les conclusions du comité Présidentiel et demande au *U.S Bureau of Census* de trouver une solution au problème. Wolter et Monsour (1981) developpent alors une procédure basée sur les filtres linéaires de la méthode X-11 (Shiskin, Young and Musgrave, 1967). Leur procédure offre une approximation simplifiée de la variance des données désaisonnalisées étant donné leurs hypothèses et le fait que les filtres linéaires eux-mêmes sont une approximation de ce que la méthode fait en pratique. Avec le même type de raisonnement Burridge et Wallis (1984) développent des modèles non-observables du type ARIMA qui approximent les filtres saisonniers du X-11 et dérivent des mesures de variance en utilisant le filtrage de Kalman (Burridge et Wallis, 1985). De même, des mesures de variance peuvent être calculés à partir des modèles ARIMA, développés par Cleveland et Tiao (1976), qui approximent le filtre symétrique de désaisonnalisation du X-11.

Ces études approchent le problème de la variance en relation avec les filtres linéaires du X-11. Cependant, ces filtres ne sont que des approximations de ce que la méthode fait en pratique sous les hypothèses de décomposition additive, d'aucun traitement des valeurs aberrantes, d'aucune variation de jours ouvrables et que seulement les options strandard sont utilisées pour estimer les composantes cyclo-tendancielle et saisonnière.

Le but de cet article est de présenter une nouvelle procédure qui approxime l'erreur quadratique moyen (EQM) des variables non-observées et de leurs changements comme elles sont estimées à partir des observations par la méthode X-11-ARIMA (Dagum, 1980), méthode utilisée par la plupart

des agences statistiques.

La section 2 de cet article introduit les modèles suppossés pour les composantes dynamiques non-observées (CDN) que sont la tendance-cycle, la saisonnalité et l'irrégulier. La section 3 donne une brève description de la méthode d'estimation pour les modèles CDN. La section 4 donne l'EQM des estimés du X-11-ARIMA et de leurs changements mensuels. La section 5 présente une application de la méthode et discute certains problèmes pratiques reliés à l'utilisation de la méthode d'approximation de la variance.

2. LA MÉTHODE X-11-ARIMA ET LES MODÈLES POUR LES COMPOSANTES NON-OBSERVÉES.

La méthode de désaisonnalisation X-11-ARIMA fait l'hypothèse que la série Y_t peut se décomposer en une composante cyclo-tendancielle C_t, saisonnière S_t et irrégulière I_t, soit de manière additive:

$$Y_t = C_t + S_t + I_t, \tag{2.1}$$

multiplicative:

$$Y_t = C_t S_t I_t \tag{2.2}$$

ou logarithmique:

$$\log Y_t = \log C_t + \log S_t + \log I_t. \tag{2.3}$$

Cette méthode est basée sur l'utilisation des moyennes mobiles ou de filtres linéaires qui implique que les composantes sont stochastiques et non pas déterministes.

Les modèles proposés pour obtenir l'EQM des données désaisonnalisées par X-11-ARIMA (pour le niveau et le changement) sont des variantes des modèles élaborés par Cleveland and Tiao (1976) et Burridge and Wallis (1984). Ces mêmes modèles furent utilisés par Kitagawa et Gersch (1984) dans leur méthode de désaisonnalisation.

Le modèle de base pour les composantes dynamiques non-observées (CDN) est de la forme:

$$Y_t = \mu_t + \gamma_t + \epsilon_t \quad , \; t=1,\ldots,T \tag{2.4}$$

où μ_t, γ_t and ϵ_t sont les composantes cyclo-tendancielle, saisonnière et irrégulière.

La composante cyclo-tendancielle est supposée suivre um modèle de la forme:

$$(1-B)^2\mu_t = \eta_t,\ \eta_t \sim N(0,\sigma_\eta^2) \qquad (2.5)$$

ou de façon équivalente:

$$\mu_t = 2\mu_{t-1} - \mu_{t-2} + \eta_t, \qquad (2.6)$$

où η_t est une séquence de variables indépendantes et identiquement distribuées (i.i.d.) et où B dénote l'opérateur de retard ($B\mu_t = \mu_{t-1}$).

Le modèle pour la composante saisonnière est défini par:

$$\gamma_t = -\sum_{j=1}^{11} \gamma_{t-j} + \omega_t,\ \omega_t \sim \text{i.i.d. } N(0,\sigma_\omega^2). \qquad (2.7)$$

Le patron de saisonnalité est donc définie par un processus qui assure que les sommes mobiles de douze composantes saisonnières ont une espérance de zéro et une variance constante à travers le temps.

Les erreurs η_t et ω_t sont indépendates entre elles et avec ϵ_t ~ i.i.d. $N(0,\sigma^2)$.

Une discussion plus détaillée sur les raisons portant sur le choix des modèles se trouve dans Dagum et Quenneville (1989).

3. ESTIMATION AVEC LES MODÈLES CDN.

L'estimation des composantes dynamiques non-observables C_t et S_t à partir des modèles (2.4), (2.5) et (2.7) se fait avec le filtrage de Kalman et le lissage sur un intervalle fixe étant donné un vecteur de conditions initiales, qui spécifie les valeurs initiales de la tendance cycle et de la saisonnalité, et les rapports de variance σ_η^2/σ^2 et σ_ω^2/σ^2.

L'estimation du vecteur des conditions initiales est obtenue à partir de la décomposition du X-11-ARIMA. Les estimateurs de vraisemblance maximale (EVM) des rapports de variance sont obtenus par la méthode du score. Soit $\underline{x}=(x_1,x_2)' = (\sigma_\eta^2/\sigma^2,\sigma_\omega^2/\sigma^2)'$ le vecteur de paramètres à être estimé. La méthode du score (Harvey, 1981) est la procédure itérative:

$$\underline{x}^{(i)} = \underline{x}^{(i-1)} + I^{-1}(\underline{x}^{(i-1)})\ \text{DlogL}(\underline{x}^{(i-1)}) \qquad (3.1)$$

avec $\underline{x}^{(i)}$ l'estimé de $\underline{x}$ à la i-ième itération, $I(\underline{x})$ est la matrice d'information de Fisher évaluée au vecteur $\underline{x}$ et $\text{DlogL}(\underline{x})$ est la matrice des dérivées du logarithme de la fonction de vraisemblance maximale évaluée à $\underline{x}$. Une description détaillée de la procédure d'estimation se trouve dans Dagum et Quenneville (1989).

4. VARIANCE DES DONNÉES DÉSAISONNALISÉES PAR X-11-ARIMA ET DE LEURS CHANGEMENTS MENSUELS.

4.1 Décomposition Additive.

Le filtrage de Kalman et le lissage sur un intervalle fixe sont appliqués à la série originale Y_t en utilisant le vecteur des conditions initiales, sa matrice de covariance, les estimés des rapports de variance σ_η^2/σ^2 et σ_ω^2/σ^2 ainsi que l'EVM de σ^2. La donnée désaisonnalisée par X-11-ARIMA est définie par $Y_t - s_t$, où s_t est l'estimé X-11-ARIMA de S_t, et la donnée désaisonnalisée des modèles CDN est définie $Y_t - \hat{\gamma}_t$, où $\hat{\gamma}_t$ est l'estimé CDN de γ_t. L'EQM est définie par:

$$EQM(\hat{\gamma}_t) = E[(Y_t-\hat{\gamma}_t)-(Y_t-\gamma_t)]^2 = E[\hat{\gamma}_t-\gamma_t]^2 \qquad (4.1)$$

et est directement obtenu à partir des matrices de covariances issues du lissage sur un intervalle fixe (Dagum et Quenneville, 1989).

Dans l'analyse des séries désaisonnalisées les comparaisons entre les mois sont faites afin d'évaluer la direction et l'importance de la tendance à court terme.

La méthode présentée ici permet l'estimation de l'EQM du changement dans les estimés désaisonnalisés des modèles CDN entre les mois t et t-a, pour a=1,...,10. Ce changement est donné par:

$$(Y_t-\hat{\gamma}_t)-(Y_{t-a}-\hat{\gamma}_{t-a}) \qquad (4.2)$$

avec EQM:

$$E[((Y_t-\hat{\gamma}_t)-(Y_{t-a}-\hat{\gamma}_{t-a}))-((Y_t-\hat{\gamma}_t)-(Y_{t-a}-\gamma_{t-a}))]^2 =$$

$$E[(\hat{\gamma}_{t-a}-\gamma_{t-a})-(\hat{\gamma}_t-\gamma_t)]^2 =$$

$$E(\hat{\gamma}_{t-a}-\gamma_{t-a})^2 + E(\hat{\gamma}_t-\gamma_t)^2 - 2E[(\hat{\gamma}_{t-a}-\gamma_{t-a})(\hat{\gamma}_t-\gamma_t)]. \qquad (4.3)$$

Encore une fois ces quantités sont disponibles à partir des matrices issues du lissage sur un intervalle fixe.

La différence standardisée entre les estimés désaisonnalisées de X-11-ARIMA et des modèles CDN est définie par:

$$D_t=(\hat{\gamma}_t-s_t)/EQM(\hat{\gamma}_t)^{1/2}. \qquad (4.4)$$

De même la différence standardisée pour le changement entre les mois t et t-a est définie par:

$$DC_t=((\hat{\gamma}_t-\hat{\gamma}_{t-a})-(s_t-s_{t-a}))/EQM(\hat{\gamma}_t-\hat{\gamma}_{t-a})^{1/2}. \qquad (4.5)$$

Sous les hypothèses de normalité, les statistiques D_t et DC_t permettent d'identifier les observations influentes lorsque les valeurs observées des statistiques sont comparées aux percentiles de la distribution normale. Pour tester l'hypothèse qu'il n'y a pas de différences significatives pour toutes les valeurs, un test du chi-deux est utilisé.

4.2 Décompositions multiplicative et logarithmique.

Dans les versions logarithmique et multiplicative les estimés des modèles CDN sont obtenus sous forme logarithmique. Ainsi pour des raisons pratiques, il est nécessaire de retransformer les données à leur forme originale. Les détails sont fournis dans Dagum et Quenneville (1989).

5. APPLICATIONS.

La désaisonnalisation des données présente des problèmes qui requièrent une attention spéciale. En particulier, la détection et le traitement des valeurs aberrantes; l'utilisation des extrapolations ARIMA afin de réduire la revision de l'ajustement contemporain; l'utilisation des facteurs contemporains ou des extrapolations prospectives des facteurs saisonniers et finalement l'ajustement direct et indirect des séries. Ces problèmes ont été pris en considération lors de l'estimation avec les modèles CDN en utilisant les mêmes procédures que X-11-ARIMA lorsqu'elles sont applicables.

La serie intitulé "Nombre total de chômeurs au Canada chez les hommes de 25 ans et plus" pour la période de janvier 1975 à décembre 1985 sert à illustrer une application typique. Une année d'extrapolation avec un modèle ARIMA$(0,1,2)(0,1,1)_{12}$ est ajouté à la série originale et la version additive est utilisée lors de la désaisonnalisation de cette série par Statistique Canada. Le tableau 1 donne les résultats de la procédure d'estimation des rapports de variance. La valeur initiale du vecteur $(\sigma_\eta^2/\sigma^2, \sigma_\omega^2/\sigma^2)$ est (.6563,.3907), la matrice des dérivées du logarithme de la fonction de vraisemblance maximale se trouve dans las deuxième colonne et la matrice d'information se trouve dans les troisième et quatrième colonnes. La valeur de log(L) (constantes exclues) est de -390.5. Finallement l'estimé de la variance des observations σ^2 est 16.855. À la sixième iteration

l'augmentation relative dans la valeur de log(L) est plus petite que .001 et la procédure est arrêtée. L'estimé final des rapports de variance ($\sigma_\eta^2/\sigma^2, \sigma_\omega^2/\sigma^2$) est (2.5605,.1151) et l'estimé final de σ^2 est 10.7422. Les valeurs des dérivées DlogL indiquent que log(L) est proche d'un maximum local à la dernière itération lorsque ces valeurs sont comparées aux valeurs à la première itération.

Étant donné les estimés des rapports de variance, les estimés des modèles CDN sont calculés et comparés avec les estimés du X-11-ARIMA. La figure 1A.1 montre la série originale et la série désaisonnalisée par X-11-ARIMA. La figure 1A.2 montre les estimés de la série désaisonnalisée par X-11-ARIMA et par les modèles CDN. La figure 1A.3 montre les différences standardisées définies par l'équation (4.4). La figure 1A.4 montre l'EQM des estimés des modèles CDN. Un test du chi-deux avec 131 degrés de liberté appliqué aux différences standardisées de la figure 1A.3 donne une valeur de 102. La probilité qu'une variable chi-deux avec 131 degrée de liberté donne une valeur observée plus grande que 102 est de .97. Ainsi l'hypothèse nulle est accepté. Dans la figure 1A.4 le graphique des EQM a une forme concave avec des sauts à toutes les années. Les EQM sont plus petites dans le milieu de la série en accord avec les résultats obtenus par Wolter et Monsour (1981) pour d'autres séries.

6. CONCLUSIONS.

Cette étude a présenté une méthode pour obtenir les EQM des données désaisonnalisées par X-11-ARIMA. Les étapes d'identification des modèles utilisées pour obtenir les EQM, d'estimation et de validation ont été brièvement présentées. La série du "Nombre total de chômeurs au Canada chez les hommes agés de 25 ans et plus" a servi à illustrer la méthode.

Pour une description plus détaillée, nous référons le lecteur à Dagum et Quenneville (1989).

BIBLIOGRAPHIE

Burridge,P. et Wallis, K.F.,(1984), "Unobserved-Components Models for Seasonal Adjustment Filters", JBES, 2, 350-359.

Burridge,P. et Wallis, K.F.,(1985), "Calculating the Variance of Seasonally Adjusted Series", JASA, 80, 541-552.

Cleveland,W.P. et Tiao,G.C.,(1976), "Decompositon of Seasonal Time Series: A Model for the Census X-11 Program", JASA, 71, 581-587.

Dagum,E.B.,(1980), "The X-11-ARIMA Seasonal Adjustment Method", Statistics Canada, Catalogue No. 12-564E.

Dagum,E.B. et Quenneville,B.,(1989), "Variance of X-11-ARIMA Seasonally Adjusted Series", TSRAD, Statistique Canada.

Harvey,A.C.,(1981), The Econometric Analysis of Time Series, Oxford, Philip Allan.

Kitagawa,G. et Gersch,W.,(1984), "A Smoothness Priors-State Space Modeling of Time Series With Trend and Seasonality", JASA, 79, 378-389.

National Commission on Employment and Unemployment Statistics,(1979), Counting the Labor Force, Washington, D.C., U.S. Government Printing Office.

President's Committee to Appraise Employment and Unemployment Statistics, (1962), Measuring Employment and Unemployment, Washington, D.C., U.S. Government Printing Office.

Shiskin,J., Young,A.H. et Musgrave,J.C.,(1967), "The X-11 Variant of the Census Method 11 Seasonal Adjustment Program", Technical Paper No. 15, U.S. Department of Commerce, Bureau of Economic Analysis.

Wolter,K.M. et Monsour,N.J.,(1981), "On the Problem of Variance Estimation for a Deseasonalized Series", Current Topics in Survey Sampling, Academic Press, 367-403.

TABLEAU 1

Nombre total de chômeurs au Canada chez les hommes de 25 ans et plus.
Resultats de la procédure d'estimation.

iter.	α	DlogL	Info1	Info2	log(L)	$\hat{\sigma}^2$
1	0.6563	23.0126	45.6511	-8.3246	-390.5	16.8555
	0.3907	-16.4130	-8.3246	73.4508		
2	1.1294	7.9834	15.7843	-3.6832	-381.4	14.2136
	0.2209	-12.6483	-3.6832	149.3000		
3	1.6183	3.2588	7.9475	-0.9249	-378.1	12.7694
	0.1483	-7.1623	-0.9249	239.9000		
4	2.0250	1.5975	5.2440	0.3890	-377.0	11.8726
	0.1200	-2.0816	0.3890	306.2000		
5	2.3302	0.9382	4.0627	0.9565	-376.6	11.2365
	0.1128	0.9849	0.9565	325.7000		
6	2.5605	0.6379	3.4278	1.1995	-376.4	10.7422
	0.1151	2.3210	1.1995	313.4000		

FIGURES 1A
NOMBRE TOTAL DE CHÔMEURS AU CANADA CHEZ LES HOMMES DE 25 ANS ET PLUS
DONNÉES DÉSAISONNALISÉES

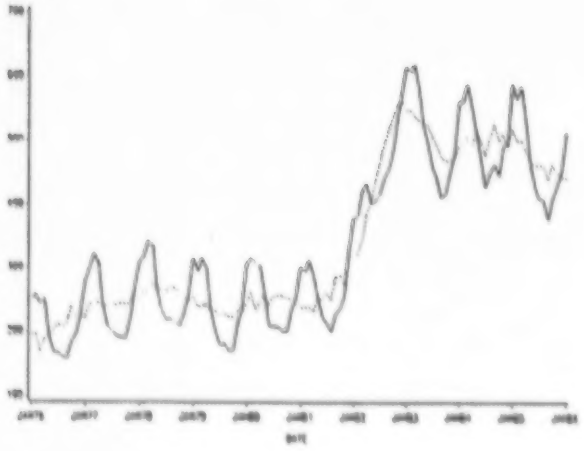


FIGURE 1A.1
___: SÉRIE ORIGINALE
...: SÉRIE DÉSAISONNALISÉE PAR X-11-ARIMA

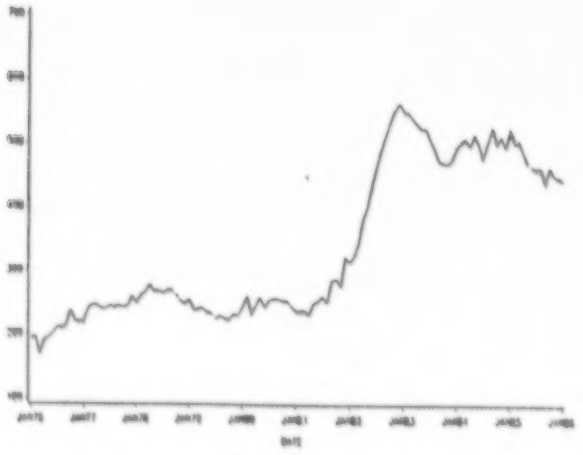


FIGURE 1A.2
___: SÉRIE DÉSAISONNALISÉE PAR X-11-ARIMA
...: SÉRIE DÉSAISONNALISÉE PAR LES MODÈLES CDN

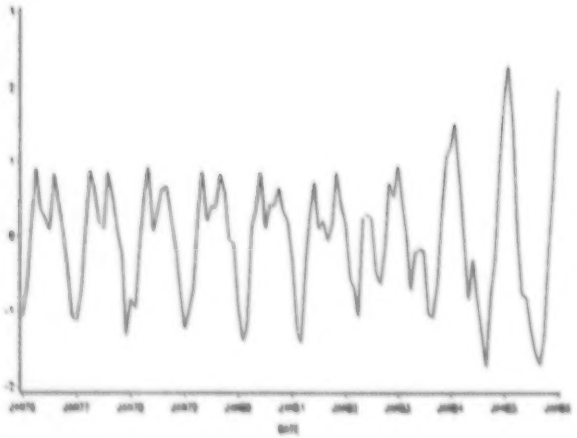


FIGURE 1A.3
___: DIFFÉRENCES STANDARDISÉES ENTRE
X-11-ARIMA ET MODÈLES CDN

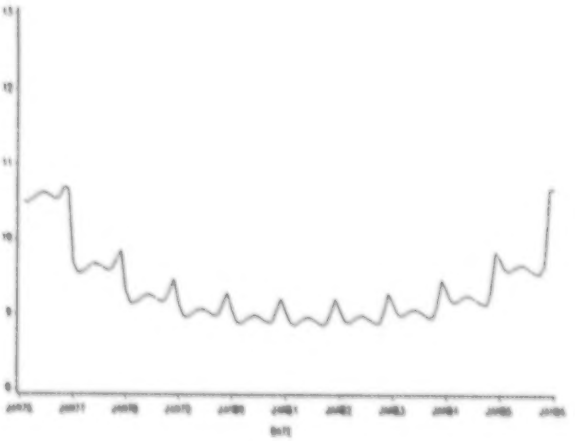


FIGURE 1A.4
___: EQM DES DONNÉES DÉSAISONNALISÉES
PAR LES MODÈLES CDN

ESTIMATION DE MODÈLES AVEC VARIABLES NON OBSERVABLES DYNAMIQUES: QUELQUES ILLUSTRATIONS UTILISANT LA MÉTHODE EM

Jacques Raynauld
Jean-Guy Simonato

Institut d'économie appliquée
École des Hautes Etudes Commerciales
5255 Decelles, Montréal, H3T 1V6

* Conférence présentée dans le cadre du Colloque «Méthodes et domaines d'application de la statistique 1989», 57e Congrès de l'A.C.F.A.S., 16 au 19 mai, Université du Québec à Montréal. Nous tenons à remercier Benoît Quenneville, l'organisateur de notre atelier.

RÉSUMÉ

De plus en plus de chercheurs modélisent les mouvements de séries chronologiques à l'aide des filtres de Kalman (voir Harvey, 1981). De nature récursive, cette méthodologie a permis l'estimation de modèles avec variables non observables dynamiques. Habituellement, les chercheurs maximisent numériquement la fonction de vraisemblance en utilisant les récursions de Kalman. Watson et Engle (1983) ont adapté la méthode EM de Dempster, Laird et Rubin (1977) au cas des modèles avec variables non observables dynamiques. La présente communication vise à clarifier les avantages de cette méthode. Deux exemples tirés de recherches actuelles serviront d'illustration.

1. INTRODUCTION

Les chercheurs ont présentement à leur disposition un éventail très riche de modèles de séries chronologiques. Les modèles *ARMA* de Box et Jenkins, *MARMA* de Tiao et Box et les méthodes de lissage sont bien connues. D'autres méthodologies plus récentes n'ont pas une utilisation aussi répandue. Le but de cette note est justement d'en illustrer une, la méthodologie des variables non observables dynamiques (aussi appelée modèles *DYMIMIC*), une approche très générale dont la formulation empirique repose sur les filtres de Kalman. Nous présentons un algorithme général dû à Dempster, Laird et Rubin (1977) et adapté aux séries chronologiques par Watson et Engle (1983) qui simplifie grandement l'estimation. Nous proposons deux applications traitant des indicateurs coincidents de la conjoncture économique québécoise et de l'évolution des primes de risque sur le marché des changes internationaux.

2. LES MODÈLES DYMIMIC

Un modèle *DYMIMIC* exprimé sous forme *state-space*[1] comprend une équation de mesure (aussi appelée équation d'observation)

$$\underset{(Nx1)}{y_t} = \underset{(NxM)}{Z}\ \underset{(Mx1)}{\alpha_t} + \underset{(Nxk1)}{\beta}\ \underset{(k1x1)}{x1_t} + \underset{(Nxn)}{S}\ \underset{(nx1)}{e_t} \tag{1}$$

et un équation de transition (aussi appelée équation d'état)

$$\underset{(Mx1)}{\alpha_t} = \underset{(MxM)}{T}\ \underset{(Mx1)}{\alpha_{t-1}} + \underset{(Mxk2)}{\delta}\ \underset{(k2x1)}{x2_t} + \underset{(Mxg)}{R}\ \underset{(gx1)}{n_t} \tag{2}$$

où e_t et n_t suivent une distribution normale centrée sur zéro avec variances H et Q respectivement. Notons que t = 1, ..., T.

L'équation de mesure (1) relie linéairement le vecteur de variables observables y_t à un vecteur de variables inobservables α_t, un vecteur de variables exogènes observables $x1_t$ et un vecteur de termes d'erreur bruit blanc e_t. Cette équation reflète l'impact de variables ou de facteurs inobservables sur les variables observables. Les matrices Z et ß sont des matrices de paramètres. La matrice S est formée de uns et de zéros. Habituellement, cette matrice est utilisée lorsqu'une des variables de y_t n'est pas contaminée par un terme d'erreur. L'équation de transition (2) spécifie l'évolution temporelle de la variable inobservable. Elle stipule le processus dynamique du vecteur α_t qui est fonction de sa valeur à la période précédente α_{t-1}, d'un vecteur de variables exogènes observables ($x2_t$) et d'un vecteur d'erreurs bruit blanc n_t. Les matrices T et δ sont des matrices de paramètres. La matrice R est formée de uns et de zéros.

Les difficultés liées à l'utilisation de tel modèles proviennent de l'estimation des T vecteurs de variables inobservables α_t et des matrices de paramètres Z, ß, T, δ, H et Q. Le problème d'estimation des T vecteurs de variables inobservables lorsque les matrices de paramètres sont connues est résolu par les filtres de Kalman et le *fixed interval smoother*. Ce dernier nous permet de trouver l'espérance de la variable inobservable conditionellement à l'information disponible de t = 0, ..., T.

---

1. La méthodologie des modèles *state-space* et des filtres de Kalman est décrite dans Harvey (1981), chapitre 4.

3. ESTIMATION DES PARAMÈTRES

Les filtres de Kalman et le *fixed interval smoother* supposent que les paramètres du modèle *state-space* sont connus. Dans la plupart des cas, les matrices de paramètres sont inconnues et doivent être estimées. Ceci peut être accompli par la méthode du maximum de vraisemblance. Pour un vecteur de paramètres θ inconnu (où θ est formé des éléments des matrices Z, ß, T, δ, H, et Q), la fonction de vraisemblance d'un modèle *state-space* peut se définir de deux façons équivalentes. En effet, le modèle décrit les observations $Y = (y_1, y_2, \ldots, y_T)'$, comme étant une fonction des variables inobservables $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_T)$ et de termes d'erreurs ayant une matrice de variance covariance V. On peut donc définir la vraisemblance comme une fonction $L(\alpha, V; \theta)$ ou comme une fonction $L(Y; \theta)$. Les deux représentations sont équivalentes car $Y = f(\alpha, V)$. Cependant, pour un modèle *state-space*, les deux représentations posent des problèmes différents en ce qui à trait à la maximisation de L par rapport à θ.

La représentation $L(Y; \theta)$, bien qu'étant facile à calculer, pose des difficultées lors de la maximisation des paramètres. C'est une fonction non-linéaire de θ donnée par:

$$(3) \qquad \ln(Y; \theta) = \text{constante} - 1/2 \sum_{t=1}^{T} (\ln |F_t| + v_t' F_t v_t)$$

où v_t et F_t représentent l'erreur de prévision et la variance de cette erreur de prévision, toutes deux obtenues des recursions des filtres de Kalman (voir Harvey 1981, page 108).

Dès que le nombre de paramètres à estimer devient important les algorithmes de maximisation deviennent coûteux, lents à converger et fastidieux à utiliser. De plus, les méthodes employées convergent souvent vers des valeurs négatives pour les paramètres des matrices de variance-covariance. Il faut alors imposer des contraintes de non-négativité ou transformer ces paramètres de façon à éviter cet inconvénient majeur. Mentionnons aussi que l'utilisation de dérivées secondes numériques lors du calcul de la matrice d'information est requise dans la plupart des cas. Ces dérivées sont souvent imprécises et très longues à calculer.

La représentation $L(\alpha, V; \theta)$ est très facile à maximiser. C'est une fonction linéaire de θ. Si le vecteur α_0 a une moyenne μ et une matrice de covariance Σ et si R et S sont des matrices identité alors $L(\alpha, V; \theta)$ est donné par:

$$\begin{aligned} (4) \qquad \ln L(\alpha, V; \theta) = {} & -1/2 \ln|\Sigma| - 1/2 (\alpha_0 - \mu)' \Sigma^{-1} (\alpha_0 - \mu) \\ & -T/2 \ln|H| - 1/2 \sum_{t=1}^{T} (\alpha_t - T\alpha_{t-1} - \delta x 2_t)' H^{-1} (\alpha_t - T\alpha_{t-1} - \delta x 2_t) \\ & -T/2 \ln|Q| - 1/2 \sum_{t=1}^{T} (y_t - Z\alpha_t - ßx1_t)' Q^{-1} (y_t - Z\alpha_t - ßx1_t). \end{aligned}$$

Les paramètres qui maximisent cette fonction sont obtenus en effectuant des OLS à partir du système d'équations formées par le modèle *state-space*. Cependant comme la matrice α est inobservable, il est impossible d'utiliser directement cette représentation. Dempster, Laird, et Rubin (1977) ont suggéré un moyen d'utiliser indirectement la représentation $L(\alpha, V; \theta)$. Il s'agit d'une procédure itérative en 2 étapes répétées jusqu'à convergence:

i) Trouver $E_i[L(\alpha, V; \theta) | Y]$ (l'espérance de $L(\alpha, V; \theta)$ conditionellement aux données observées Y et aux paramètres θ_i),
ii) Trouver θ_{i+1} qui maximise $E_i(L)$.

Cet algorithme, appelé EM pour *Expectation-Maximisation*, a été adapté par Shumway et Stoffer (1982) et Engel et Watson (1983) aux modèles *DYMIMIC* exprimés sous forme *state-space*. A la première étape, $E(L(\alpha,V;\theta)|Y)$ est facilement calculé à l'aide des filtres de Kalman et du *fixed interval smoother*. La seconde étape effectue des OLS à partir des valeurs calculées à la première étape. L'algorithme ne requiert aucune dérivée. De plus Dempster, Laird et Rubin ont montré que, sous certaines conditions de régularité, l'algorithme augmente à chaque itération la valeur de la fonction de vraisemblance jusqu'à l'obtention d'un maximum local. Nous avons programmé cet algorithme en langage GAUSS 2,0. Notons que la méthode EM n'associe aucun écart-type aux estimés obtenus. Afin d'obtenir cette information, nous avons procédé à quelques itérations par la méthode du score (voir Engle et Watson, 1982) en utilisant les estimés de la méthode EM comme valeurs de départ.

4. PREMIER EXEMPLE: INDICATEUR COÏNCIDENT DE L'ÉCONOMIE QUÉBÉCOISE.

Il n'existe pas présentement d'indicateur mensuel fiable de l'activité économique au Québec. Les mesures agrégées canadiennes telles l'indice du produit intérieur brut ou le produit domestic brut ne sont généralement pas disponibles sur une base mensuelle pour les entités provinciales. Pourtant, il est essentiel d'avoir une mesure coincidente des performances de l'économie. Stock et Watson (1988) ont montré que la méthodologie *DYMIMIC* est une solution à ce problème. Supposons que l'indicateur idéal de l'activité économique québécoise existe mais qu'il ne soit pas observable. Notons cet indicateur C_t. Seules les trois variables livraisons, ventes et emplois sont connues à chaque période[2]. Prises séparément, ces trois variables sont des indicateurs imparfaits de l'activité économique globale, chacune traçant un portrait de son secteur propre. Cependant, elles peuvent nous renseigner sur l'évolution de la variable non observable puisqu'elles y sont directement reliées.

Plus spécifiquement, nous avons postulé et estimé trois équations d'observation où chaque variable observable est décomposée en une partie commune et une composante propre;

$$(5)\quad (1-L^{12})\,\text{Livraisons}_t = \underset{(4,64)}{0,2248}\,(1-L^{12})C_t + u1_t$$

$$(6)\quad (1-L^{12})\text{ventes}_t = \underset{(7,66)}{0,3210}\,(1-L^{12})C_t + u2t$$

$$(7)\quad (1-L^{12})\text{emplois}_t = \underset{(4,27)}{0,.1963}\,(1-L^{12})C_t + u3_t$$

et les quatre équations de transition pour les variables non observables commune et spécifiques

$$(8)\quad (1-L^{12})C_t = \underset{(32,0)}{0,9351}\,(1-L^{12})C_{t-1} + vt \qquad \hat{\partial}_v = 1,0$$

$$(9)\quad u1_t = \underset{(14,7)}{0,7429}\,u1_{t-1} + e1t \qquad \hat{\partial}_{e1} = 0,56$$

$$(10)\quad u2t = \underset{(0,04)}{0,0055}\,u2_{t-1} + e2_t \qquad \hat{\partial}_{e2} = 0,43$$

$$(11)\quad u3_t = \underset{(16,8)}{0,7914}\,u3_t\text{-}1 + e3t \qquad \hat{\partial}_{e2} = 0,47$$

---

2. Ces trois variables sont les plus représentatives des indicateurs mensuels disponibles pour le Québec.

La variance associée au terme d'erreur de la variable inobservable a été fixée à 1.0 pour fin d'identification et toutes les series ont subi une transformation logarithmique. La matrice de variance covariance de l'équation de transition est diagonale puisque les covariances sont postulées nulles. $(1-L^{12})C_t$ représente le taux de croissance annuel de «l'état de l'économie québécoise» et ui_t représente la portion individuelle non observable propre à chaque série; ces quatre variables sont modélisées à l'aide d'un AR(1)[3]. Livraisons représente les livraisons des industries manufacturières au Québec (série # D319626); ventes sont les estimations mensuelles des ventes de commerces de détail au Québec (D650555) et emploi correspond aux emplois au Québec (D768414). Les ventes ont été déflatées par l'indice des prix à la consommation (D486056) et les livraisons par l'indice des prix des ventes dans l'industrie manufacturière (D500000 et D614001). Toutes les series ont été ajustées pour la saisonnalité par la procédure XARMA-11 et ont toute été centrées et réduites. Les tests t habituels sont donnés sous les estimés des paramètres. Rappelons que ces tests sont calculés après quelques itérations par la méthode du score.

Le modèle estimé indique que les trois variables observées sont modérément reliées à une variable non observable commune représentant l'activité économique globale. Cette composante commune est relativement lisse avec un coefficient autorégressif de 0,93 (voir graphique 1 qui montre l'évolution de chaque série et de la composante commune associée). Les composantes propres à chaque série sont plus erratiques; c'est le cas des ventes dont le coefficient est faible et non significatif. Une modélisation plus approfondie devrait être effectuée puisque nous avons détecté des problèmes d'autocorrélation de délai 12.

5. DEUXIÈME EXEMPLE: PRIME DE RISQUE SUR LES TAUX DE CHANGE

De nombreuses recherches ont tenté d'identifier une prime de risque sur les marchés de change (voir Wolff, 1987). Cette prime de risque doit être extraite de l'erreur de prévision définie comme la différence entre les taux de change au comptant et à terme. Si la prime de risque est inexistante, l'erreur de prévision devrait se comporter comme un bruit blanc. Plus spécifiquement, nous cherchons à estimer une prime de risque commune entre trois monnaies (franc français, franc suisse et le mark allemand) sous l'hypothèse de marchés efficaces. La modélisation prend la forme de trois équations d'observations

$$(12) \quad eff_t = \underset{(13,8)}{0,9795\ p_t} + e1_t \qquad \hat{\sigma}_1 = 0,01$$

$$(13) \quad efs_t = \underset{(13,2)}{1,0866\ p_t} + e2_t \qquad \hat{\sigma}_2 = 0,01$$

$$(14) \quad ema_t = \underset{(9,04)}{0,7244\ p_t} + e3_t \qquad \hat{\sigma}_3 = 0,02$$

et d'une équation de transition de type AR(1) pour la prime de risque non observable

$$(15) \quad p_t = \underset{(-0,24)}{-0,022\ p_{t-1}} + v_t \qquad \hat{\sigma}^2_v = .001$$

---

3. Initialement, nous voulions modéliser la première différence des trois séries puisque des tests de stationnarité indiquaient la présence d'une racine unitaire. Cependant, l'analyse de telles séries transformées était difficile puisqu'elles s'apparentaient à des bruits blancs. Pour contrer ce problème, nous avons effectué une différence douzième pour investiguer les changements annuels, changements plus appropriés à une analyse comm la nôtre.

eff_t est la différence entre le logarithme des taux de change au comptant et à terme pour le franc français; efs_t correspond au franc suisse et ema_t au mark allemand. p_t est la prime de risque non observable qui suit une structure AR(1).

Les résultats montrent qu'il n'est pas possible d'extraire une prime de risque commune qui a une structure AR(1) puisque la composante pt est une bruit blanc (coefficient faible et non significatif, voir aussi le graphique 2 qui montre l'évolution de la prime de risque commune relative à chaque monnaie). Cependant, une analyse approfondie a montré que les termes d'erreur ne sont pas distribués comme un bruit blanc ce qui indique l'existence d'une prime de risque propre à chaque monnaie. Idéalement, la modélisation devrait être poursuivie en adoptant une stratégie similaire au cas des indicateurs.

6. CONCLUSION

Avec l'algorithme d'estimation EM, la modélisation *DYMIMIC* devrait être plus accessible comme outil de modélisation de séries chronologiques. C'est ce que nous espérons avoir illustré. Dans la plupart des cas, la convergence se fait rapidement et à peu de frais puisque nous l'avons implanté à l'aide du logiciel GAUSS 2,0 disponible sur ordinateur de type IBM PC. Une fois les estimés obtenus, les tests t sont calculés en quelques itérations par la méthode du score. Les deux exemples choisis n'ont pas été traités de façon exhaustive, l'exercice visant plutôt à illustrer les possibilités d'application.

BIBLIOGRAPHIE

Dempster, A.P., N.M. Laird, and D.B. Rubin (1977), "Maximum likelihood from incomplete data via the EM algorithm", *J. Roy. Stat. Soc., Ser.B.* 39, 1-38.

Harvey, A.C. (1981), *Time Series Analysis*, Halsted Press.

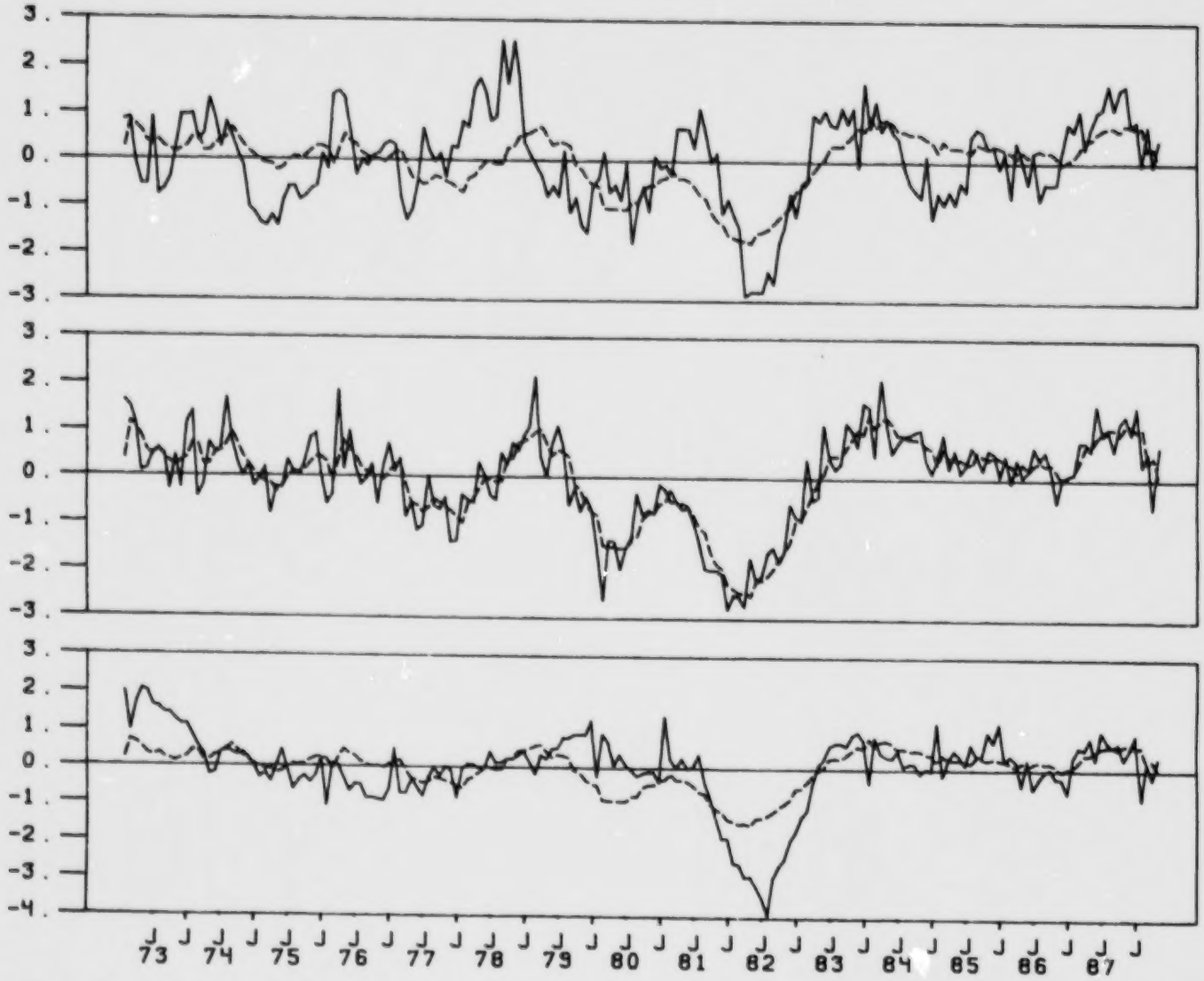
Shumway, R.H., and D.S. Stoffer (1982), "An approach to time series smoothing and forecasting using the EM algorithm", *J. Time Series Anal.* 3, 253-264.

Stock, J.H. and M.W. Watson, (1988), "A Probability Model of the Coincident Economic Indicators", NBER Discussion Paper No. 2772.

Watson, M.W. and R.F. Engle (1983), "Alternative Algorithms for Estimation of Dynamic MIMIC, Factor, and Time Varying Coefficient Regression Models", *Journal of Econometrics* 23, 385-400.

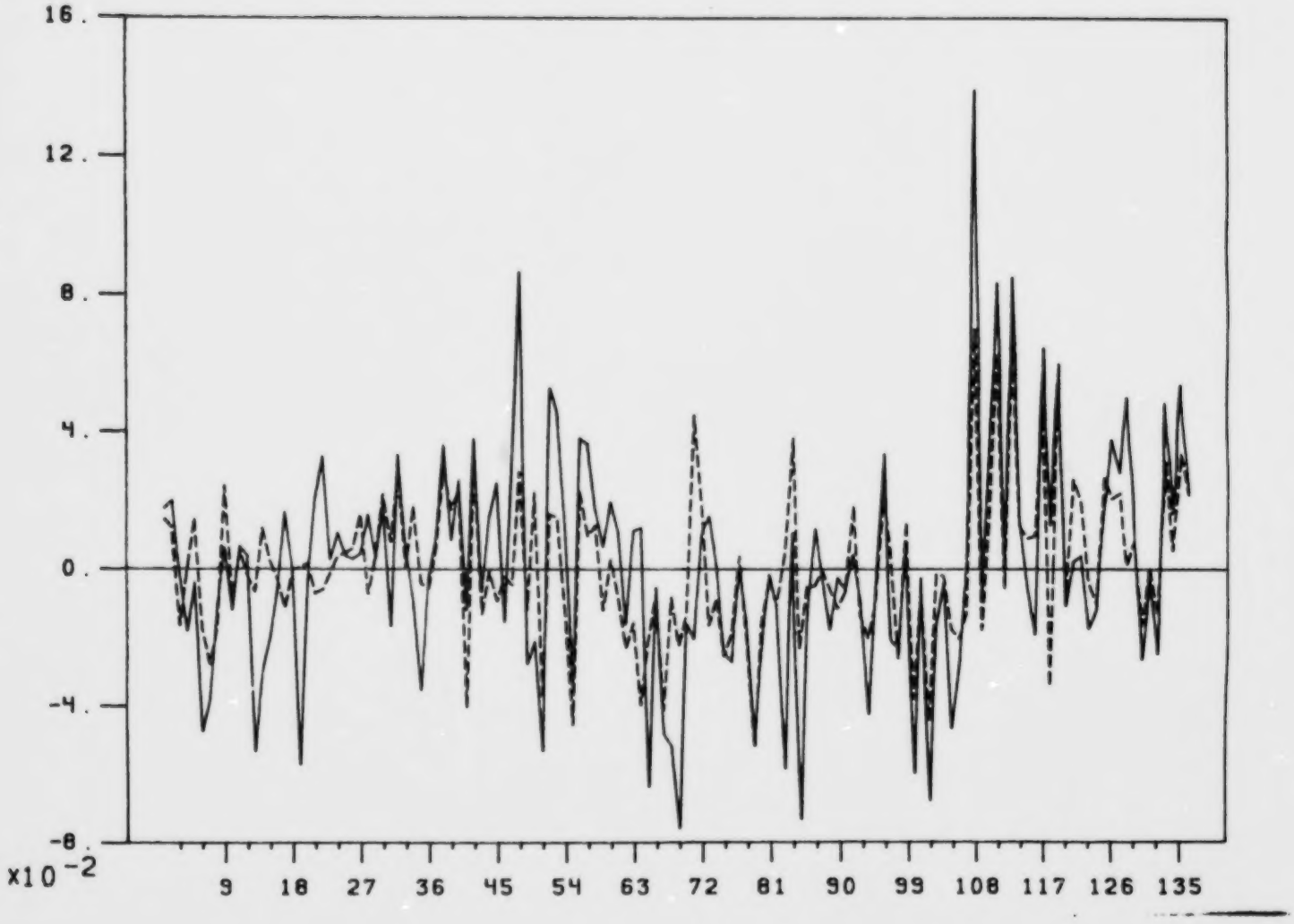
Wolff, Christian C.P. (1987), "Forward Foreign Exchange Rates, Expected Spot Rates, and Premia: a Signal Extraction Approach", *Journal of Finance* 42, No. 2, 395-406.

GRAPHIQUE 1: INDICATEUR COÏNCIDENT DU QUÉBEC



Les graphiques illustrent le comportement des livraisons, ventes et emploi. Les lignes pointillées correspondent à la partie commune C_t ajustée par chacun des coefficients des équations d'observations.

GRAPHIQUE 2: PRIME DE RISQUE



Le graphique illustre le comportement des erreurs de prévision du deutch mark. La ligne pointillée correspond à la partie commune C_t ajustée par le coefficient de l'équation d'observation.

SESSION SUR LES PLANS D'EXPÉRIENCE

MÉTHODES POUR TRAITER L'INTERACTION DANS UN PLAN D'ÉTUDE D'UN ESSAI CLINIQUE

PAR: YVES LALONDE

(LABORATOIRES NORDIC)

2775, Autoroute des Laurentides

Laval, QC, Canada

RÉSUMÉ

Le présent article propose deux façons de traiter une interaction significative à l'intérieur d'un plan d'expérience en recherche clinique;

A) Cas I: Interaction traitement-par-visite traité avec l'analyse en composante principale

B) Cas II: Interaction traitement-par-centre; pourcentage de contribution

Ces méthodes se veulent des techniques exploratrices dans le but de faciliter l'analyse et l'interprétation des données.

L'application de ces méthodes peut être adaptée pour répondre à d'autres besoins spécifiques.

MÉTHODES POUR TRAITER L'INTERACTION DANS UN PLAN D'UN ESSAI CLINIQUE

A) CAS I

INTRODUCTION

En recherche clinique, il est commun de mesurer un même paramètre à différents intervalles de temps. Il en résulte souvent des observations qui sont fortement corrélées entre elles. Pour fin comparative, le modèle paramétrique le plus souvent utilisé est le plan factoriel en mesures répétées (split splot design [1]) avec facteurs principaux les visites d'évaluation et les traitements administrés.

		VISITE 1	VISITE 2	• • •	VISITE k
TRAITEMENT	A	Y	Y	• • •	Y
	B	Y	Y	• • •	Y

Le modèle linéaire se présente sous la forme suivante:

$$Yijk = \mu + \alpha i + \pi k(i) + \beta j + \alpha\beta ij + \beta\pi jk(i) + \Sigma ijk,$$

où $i = 1,\ldots, p;\ J = 1,\ldots, q;\ k = 1,\ldots, ni;$

ET μ = moyenne globale

αi = l'effet du $i^{ème}$ traitement

$\pi k(i)$ = k ème sujet à l'intérieur du $i^{ème}$ traitement

βj = J ème visite

$\alpha\beta ij$ = l'interaction traitement-par-visite

$\beta\pi jk(i)$ = l'interaction patient-par-visite (erreur A)

Σijk = Erreur B

Nous assumons que les hypothèses sous-jacentes au modèle sont respectées.

L'application de ce modèle ne permet cependant pas de déterminer des intervalles de temps critiques s'ils existent. De plus, si une interaction traitement-par-temps est décelée, la stratégie afin d'identifier les différences entre traitement n'est pas évidente. De plus, l'approche multivariée en mesures répétées n'est valide qu'en l'absence d'interaction significative.

MÉTHODOLOGIE

En recherche clinique, certains paramètres sont mesurés à plusieurs reprises i.e. suivant de longs intervalles de temps.

Nous avons voulu identifier une méthode objective afin de réduire la dimension des observations mesurées dans le temps. Nous cherchions un "ré-arrangement" contigu des observations dans le temps (que l'on définira éventuellement par "cycle"). Enfin, puisque les données originales sont fortement corrélées entre elles, nous cherchions des regroupements (préférablement) indépendants.

Afin d'identifier des cycles contigus, la méthodologie suivante est proposée:

1- Application de l'analyse en composante principale sur les valeurs d'un paramètre à différentes visites et où le calcul des valeurs et vecteurs propres est exécuté de façon usuelle [2].

2- Seules les composantes principales ayant une valeur propre supérieure à 1 sont retenues.

3- Application d'une rotation varimax afin d'identifier des cycles (regroupement de visites contiguës) ayant le plus de poids à l'intérieur de chaque vecteur propre.

4- Si la rotation ne permet pas d'associer à chaque cycle un ensemble de visites contiguës, une rotation oblique est alors exécutée (selon notre expérience, cette rotation oblique s'est souvent avérée efficace).

5- Création de "nouvelles variables" qui correspondent à la moyenne des observations qui sont le plus fortement corrélées à l'intérieur de chaque composante principale.

6- Application d'un plan factoriel en mesures répétées sur chacun des cycles.

ILLUSTRATION

Nous allons illustrer la méthodologie présentée à l'aide d'un exemple. La variable étudiée est la pulsation cardiaque mesurée selon 2 traitements A ou B, à 15 intervalles de temps (voir Figure 1).

Les résultats de l'analyse de variance en mesures répétées (SAS GLM [3]) sont présentés dans le Tableau 1. Une interaction significative traitement-par-visite ($p < 0.0001$) y est décelée.

L'analyse en composante principale (après rotation oblique) fournit quatre facteurs ayant une valeur propre supérieure à l'unité, expliquant 78% de la variation totale (voir Tableau 2). Ici, l'arrangement des facteurs est fonction de l'ordre de grandeur des valeurs propres associée à chaque facteur et est sans importance dans

notre analyse. Nous considérons tous les quatres facteurs ayant une importance équivalente.

Pour chaque facteur, on identifie les cycles ayant le plus grand coefficient à l'intérieur de chaque vecteur propre. Ainsi, il en résulte que le premier facteur inclut les cycles 3 à 8; le suivant inclut les cycles 9 à 12; le troisième inclut les cycles 13 à 15; et le dernier les cycles 1 et 2 (voir Tableau 2).

L'étape suivante consiste à former selon chaque traitement les valeurs moyennes correspondant aux cycles appropriés. Le Tableau 3 illustre ces valeurs moyennes ainsi que les valeurs-p résultant du modèle linéaire. Une représentation graphique est donnée à la Figure 2 pour chacun des traitements en relation avec les cycles définis.

A l'intérieur de chacune des quatres analyses, il n'y a pas d'interaction significative et le modèle linéaire nous permet de conclure qu'il y a une différence significative entre les deux traitements en début d'étude (jusqu'à la Visite 8) (voir Tableau 3).

CONCLUSION

Afin de faciliter l'analyse en mesures répétées, l'A.C.P. s'avère un outil efficace dans le but de regrouper les observations en cycles contigus.

L'échantillon résultant produit des valeurs quasi indépendantes ainsi qu'une analyse plus sensible servant à identifier les différences entre traitements dans le temps.

Cependant, la méthodologie n'est pas toujours applicable selon:

1) que la variation expliquée par l'A.C.P. n'est pas suffisante;
2) que l'application de l'A.C.P. ne permet pas une réduction suffisante du système;
3) qu'à l'intérieur de chaque facteur, il n'est pas possible de cerner des cycles contigus.

---

[1]:Winer, B.J., Statistical Principles in Experimental Design. New York, McGraw Hill, 2nd Edition, 1971.

[2]:Kendall, M.G., A course in Multivariate Analysis, Charles Griffin and Company Ltd, London, 1961.

[3]:S.A.S. User's Guide: Basics and Statistics, SAS Institute Inc. 1985 Edition.

FIGURE 1

REPRESENTATION GRAPHIQUE DES OBSERVATIONS

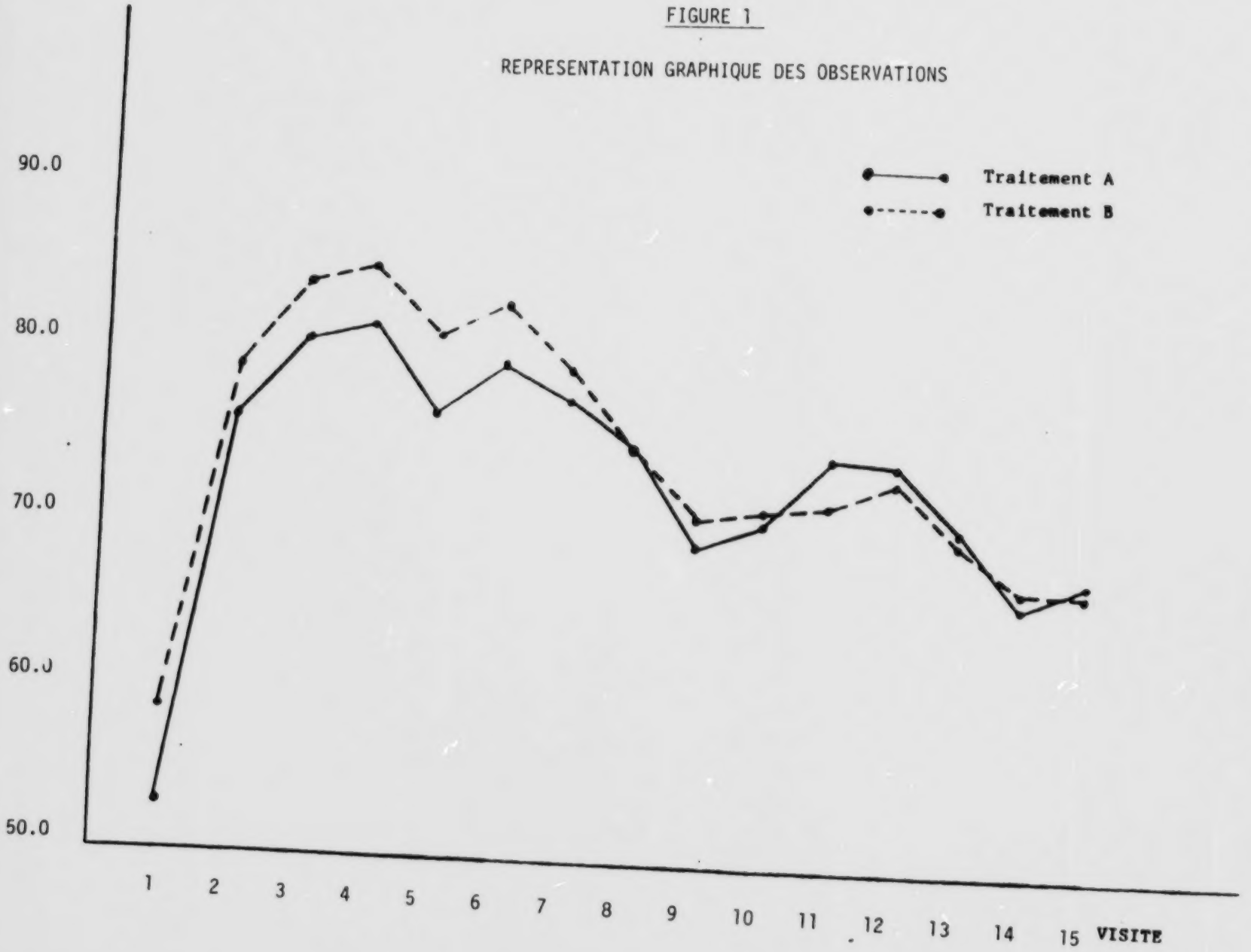


TABLEAU 1

RÉSULTATS DU PLAN FACTORIEL (2x15) EN MESURES RÉPÉTÉES

(SAS GLM)

Source	D.L.	Somme des carrés	F
Traitement	1	1686	3.02 (P<.09)
Erreur (A)	46	559	
Visite	14	6578	126.82 (P<.0001)
Traitement*Visite	14	1653	3.18 (P<.0001)
Erreur (B)	1844	52	

TABLEAU 2

ANALYSE EN COMPOSANTE PRINCIPALE

(APRÈS UNE ROTATION DE TYPE OBLIQUE)

Visite	Facteur 1	Facteur 2	Facteur 3	Facteur 4
1	0.24	0.12	0.18	0.86
2	0.64	0.24	0.09	0.82
3	0.75	0.23	0.09	0.70
4	0.83	0.17	0.09	0.59
5	0.88	0.21	0.17	0.38
6	0.91	0.31	0.19	0.34
7	0.83	0.48	0.15	0.29
8	0.78	0.49	0.24	0.13
9	0.63	0.74	0.22	0.09
10	0.35	0.89	0.25	0.10
11	0.24	0.91	0.46	0.20
12	0.29	0.74	0.63	0.20
13	0.12	0.57	0.82	0.14
14	0.17	0.36	0.93	0.16
15	0.20	0.26	0.86	0.07
	Cycles 3-8	Cycles 9-12	Cycles 13-15	Cycles 1-2

FIGURE 2

REPRESENTATION GRAPHIQUE DES VALEURS MOYENNES DES CYCLES

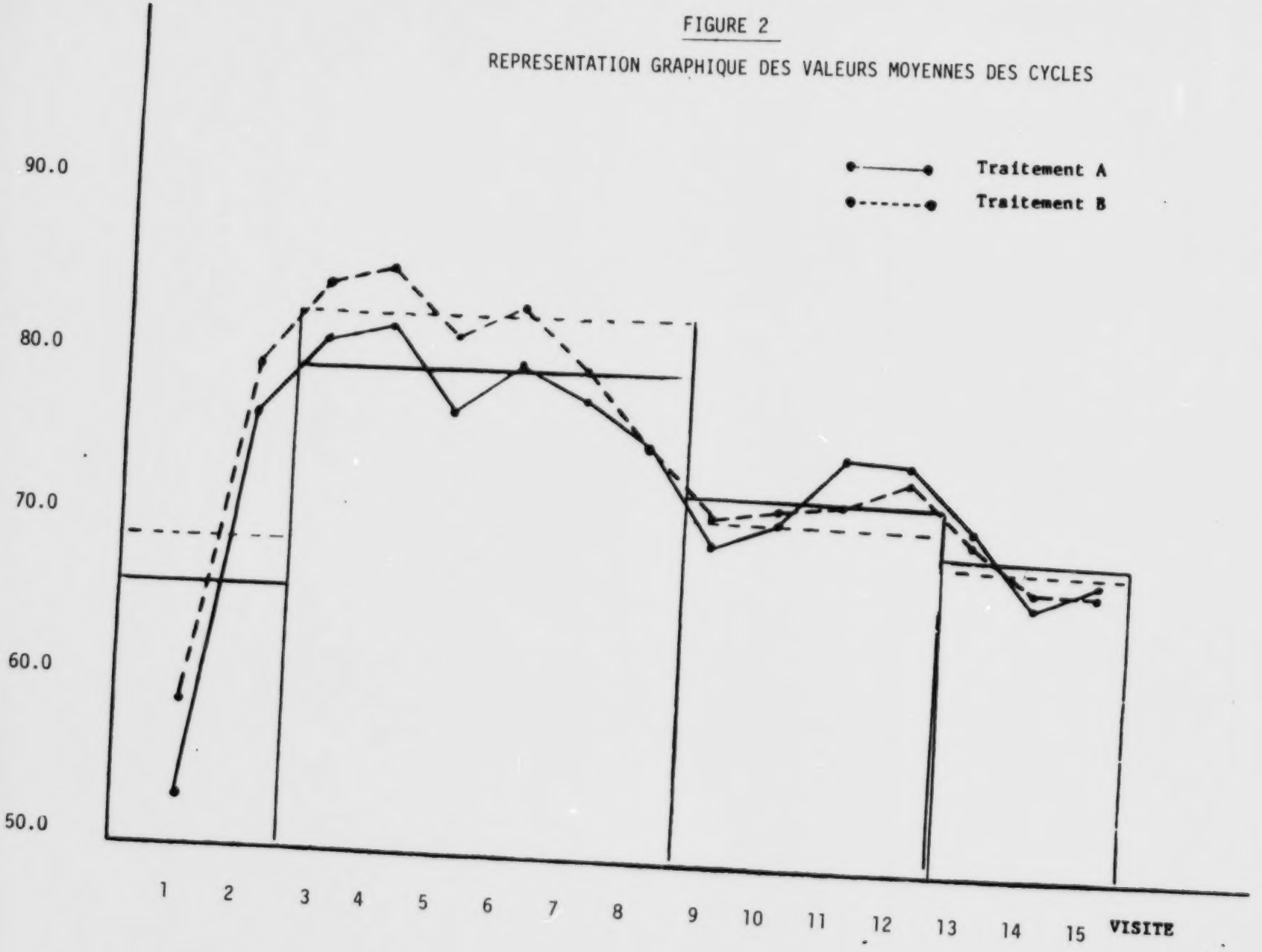


TABLEAU 3

ANALYSE PAR CYCLE

	Traitement A	Traitement B	Comparaison
Cycles 1-2	68.6	64.7	(P<.01)
Cycles 3-8	81.8	78.4	(P<.01)
Cycles 9-12	71.9	72.3	(NS)
Cycles 13-15	68.1	68.2	(NS)

MÉTHODES POUR TRAITER L'INTERACTION DANS UN PLAN D'ÉTUDE D'UN ESSAI CLINIQUE

B)- CAS II

INTRODUCTION

Dans un essai clinique multicentrique, l'examen de l'interaction globale d'un plan factoriel IXJ est une étape importante afin de justifier de réunir les observations provenant de chacun des sites.

Le modèle factoriel IXJ se définit comme suit:

$$Yijk = \mu + \alpha i + \beta j + \alpha\beta ij + \Sigma ijk,$$

où μ = moyenne globale

αi = l'effet du $i^{ème}$ traitement

βj = l'effet du $j^{ème}$ centre

$\alpha\beta ij$ = l'interaction traitement-par-centre

Si l'interaction traitement-par-centre s'avère significative, les principales questions qui surviennent sont:

A) Quel(s) investigateur(s) est (sont) responsable(s) de l'interaction?

B) Quel est le pourcentage de contribution apportée par chacun des centres?

MÉTHODOLOGIE

A) Afin de déterminer quels centres sont différents des autres, il suffit de générer des contrastes pairés (interaction pairée) comparant les résultats de chaque centre avec tous les autres.

Dans le cas de 2 traitements, l'interaction pairée entre le centre i et j se définit de la façon suivante:

$$tij = \frac{[\bar{Y}1i - \bar{Y}2i] - [\bar{Y}1j - \bar{Y}2j]}{s\,[\,1/_{n1i} + 1/_{n2i} + 1/_{n1j} + 1/_{n2j}]^{\frac{1}{2}}}$$

où tij est une statistique -t possédant une erreur s et un nombre de degré de liberté associé au modèle d'analyse de variance IXJ.

Cette étape nous permet d'identifier combien de fois un centre donné "interacte" avec tous les autres.

Ceux dont la fréquence est plus élevée sont souvent ceux qui contribuent le plus à l'interaction globale.

Pour fin de comparaison, le seuil de signification est fixé à 5%.

- Des comparaisons "non-multiples" sont générées ne nécessitant pas d'ajustement du seuil de signification.

- Un ajustement du type boneferroni pour comparaisons multiples peut être appliqué au choix.

- L'énoncé "contrast" de la procédure GLM (logiciel SAS) réussit à générer les contrastes appropriés considérant que le nombre de centre participant à un essai clinique varie d'une variable à l'autre et/ou d'une visite à l'autre.

B) POURCENTAGE DE CONTRIBUTION

Lorsqu'un centre a été identifié comme étant responsable d'une interaction globale significative, la prochaine question qui se pose est de savoir jusqu'à quel point est-il responsable de la somme des carrés totale pour l'interaction.

C'est ce que nous référons au "pourcentage de contribution".

Dans le cas de 2 traitements, les calculs s'effectuent de la façon suivante:

SOIT: $n1j$ = taille d'échantillon du centre J pour le traitement 1

$n2j$ = taille d'échantillon du centre J pour le traitement 2

$\bar{Y}1j$ = valeurs moyennes du centre J pour le traitement 1

$\bar{Y}2j$ = valeurs moyennes du centre J pour le traitement 2

ALORS $Dj = (\bar{Y}1J - \bar{Y}2j)$ est la différence entre les 2 moyennes pour le centre J;

$$\bar{D} = \sum_j Wj\ Dj\ /\ \sum_j Wj$$

où $Wj = n1j\ n2j\ /\ (n1j + n2j)$

est la moyenne globale pondérée provenant de tous les centres;

$Cj = \bar{W}j\ [Dj - D]^2$ est l'écart du centre J;

$$PCj = Cj\ /\ 100\ \sum_j CJ$$

est le pourcentage de contribution du centre J à la somme des carrés de l'interaction.

- Des calculs similaires peuvent être exécutés dans le cas de plus de 2 traitements.

- Les valeurs peuvent être obtenues par sommation parmi tous les investigateurs, des carrés des différences entre les valeurs prédites d'un modèle sans interaction du modèle avec interaction. (Options output et predicted de SAS/GLM en calculant les sommes et différences dans l'étape data).

ILLUSTRATION

Un paramètre (Y1) a été mesuré à 2 reprises selon 4 différents modes de traitement (# 149, 163, 190 et 216) à travers 5 investigateurs (block) pour constituer un échantillon total de N=40 observations.
Les valeurs moyennes de chaque traitement sont illustrées à la Figure 1 pour chaque investigateur. Il apparaît clairement que l'investigateur #1 est la source principale d'une interaction traitement-par-investigateur significative. En effet, suivant le modèle linéaire présenté dans le Tableau 1, une interaction globale est décelée ($p = 0.0004$).

Contribution à l'interaction: Les contrastes pairés entre les investigateurs montrent que le Centre 1 interacte de façon significative avec les quatres autres (voir Tableau 1).

Pourcentage de contribution: Le pourcentage de contribution est calculé pour chacun des investigateurs montrant une contribution à l'interaction de l'ordre de près de 73% pour l'investigateur #1 (voir Tableau 2).

FIGURE 1

CONTRIBUTION A L'INTERACTION
(PAR CELLULE)

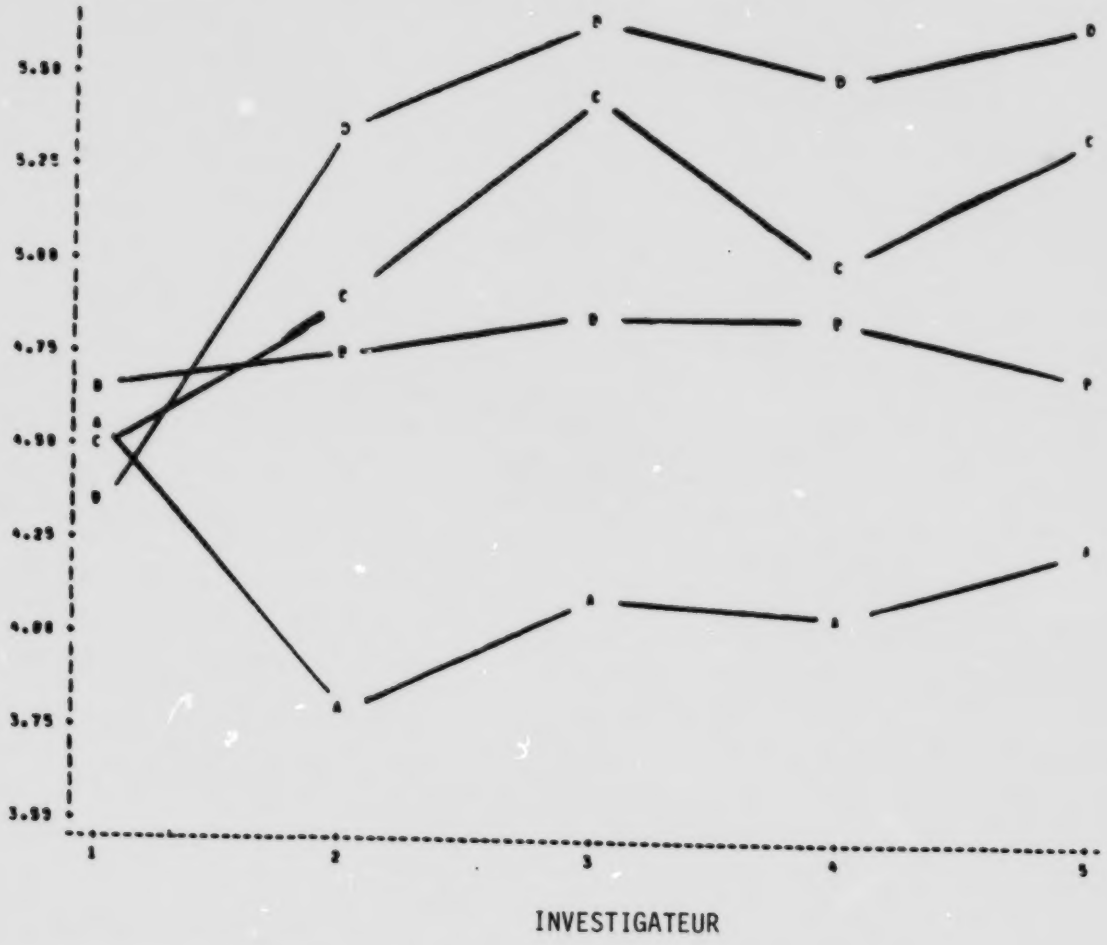


TABLEAU 1

CONTRIBUTION A L'INTERACTION

GENERAL LINEAR MODELS PROCEDURE

DEPENDENT VARIABLE: Y

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	PR > F
MODEL	19	11.58482408	0.60972758	15.23	0.0001
ERROR	20	0.80054913	0.04002746		
CORRECTED TOTAL	39	12.38537320			

R-SQUARE	C.V.	ROOT MSE	Y MEAN
0.935363	4.1558	0.20006863	4.81424952

SOURCE	DF	TYPE I SS	F VALUE	PR > F
TRT	3	7.52952493	62.78	0.0001
BLK	4	1.35641486	8.47	0.0004
TRT*BLK	12	2.69888428	5.62	0.0004

SOURCE	DF	TYPE III SS	F VALUE	PR > F
TRT	3	7.52952493	62.78	0.0001
BLK	4	1.35641486	8.47	0.0004
TRT*BLK	12	2.69888428	5.62	0.0004

CONTRAST	DF	SS	F VALUE	PR > F
INT. BL 1 - BL 2	3	1.56656814	13.84	0.0001
INT. BL 1 - BL 3	3	1.82471820	15.20	0.0001
INT. BL 1 - BL 4	3	1.33846820	11.15	0.0002
INT. BL 1 - BL 5	3	1.51164977	12.59	0.0001
INT. BL 2 - BL 3	3	0.08664999	0.72	0.5508
INT. BL 2 - BL 4	3	0.01514996	0.13	0.9435
INT. BL 2 - BL 5	3	0.13381088	1.11	0.3667
INT. BL 3 - BL 4	3	0.11369992	0.95	0.4360
INT. BL 3 - BL 5	3	0.04881875	0.41	0.7500
INT. BL 4 - BL 5	3	0.10786889	0.90	0.4594

TABLEAU 2

STATISTIQUES PAR CELLULE
CONTRIBUTION A L'INTERACTION

VARIABLE= Y1 BLOCK=INV1

TREAT-MENT	CELL SIZE	CELL MEAN	CELL STD	CONT-RIBUTION	% CONT-RIBUTION	% OF PATIENTS
149	2	4.50	0.66	0.9681	35.87	5.00
163	2	4.63	0.08	0.0610	2.29	5.00
196	2	4.52	0.10	0.1370	5.06	5.00
216	2	4.37	0.16	0.0198	30.36	5.00
---				------	-----	-----
BLK				1.9567	72.58	20.00

VARIABLE= Y1 BLOCK= INV2

TREAT-MENT	CELL SIZE	CELL MEAN	CELL STD	CONT-RIBUTION	% CONT-RIBUTION	% OF PATIENTS
149	2	3.88	0.89	0.1189	4.11	5.00
163	2	4.73	0.25	0.0160	0.59	5.00
196	2	4.92	0.25	0.0006	0.02	5.00
218	2	5.37	0.03	0.0535	1.98	5.00
---				------	----	-----
BLK				0.1910	6.71	20.00

VARIABLE= Y1 BLOCK= INV3

TREAT-MENT	CELL SIZE	CELL MEAN	CELL STD	CONT-RIBUTION	% CONT-RIBUTION	% OF PATIENTS
149	2	4.08	0.07	0.1386	4.94	5.00
163	2	4.83	0.04	0.0244	0.90	5.00
196	2	5.43	0.19	0.6798	2.89	5.00
218	2	5.67	0.01	0.0569	2.10	5.00
---				------	-----	-----
BLK				0.2898	10.74	20.00

VARIABLE= Y1 BLOCK= INV4

TREAT-MENT	CELL SIZE	CELL MEAN	CELL STD	CONT-RIBUTION	% CONT-RIBUTION	% OF PATIENTS
149	2	4.04	0.09	0.0341	1.26	5.00
163	2	4.83	0.06	0.0059	0.22	5.00
196	2	5.01	0.01	0.0091	0.34	5.00
218	2	5.48	0.26	0.0412	1.53	5.00
---				------	----	-----
BLK				0.0903	3.35	20.00

VARIABLE= Y1 BLOCK= INV5

TREAT-MENT	CELL SIZE	CELL MEAN	CELL STD	CONT-RIBUTION	% CONT-RIBUTION	% OF PATIENTS
149	2	4.25	0.13	0.0110	0.41	5.00
163	2	4.72	0.11	0.0876	3.24	5.00
196	2	5.35	0.02	0.0282	1.04	5.00
218	2	5.66	0.20	0.0543	2.01	5.00
----				------	------	------
BLK				0.1811	6.71	20.00
IVAR				2.6989	100.00	100.00
				******	******	******
				2.6989	100.00	100.00

Détermination des causes d'orniérage à l'aide de la planification d'expérience (suite et fin)

Mario Montégiani*

RÉSUMÉ

Cette communication se veut une suite de celle présentée lors du dernier colloque sur les méthodes et domaines d'application de la statistique dans le cadre du 56e congrès de l'ACFAS qui se déroulait à Moncton en 1989. Ainsi, on présentera les différentes étapes qui suivent généralement une planification: cueillette des données, analyses, résultats et tous les problèmes qui se rattachent à une expérience de ce genre.

* Mario Montégiani est statisticien au Service de la statistique et de l'économie des transports au Ministère des Transport du Québec.

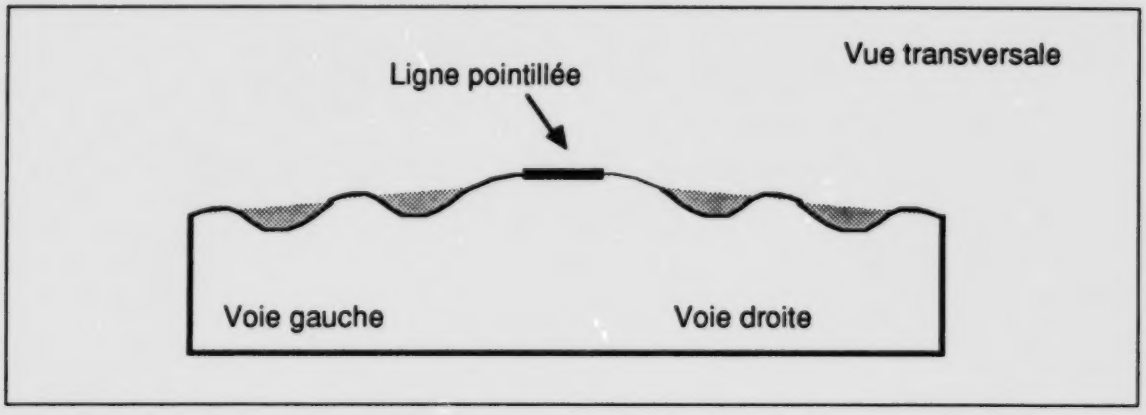
Détermination des causes d'orniérage à l'aide de la planification d'expérience (suite et fin)

L'orniérage est le problème des années 80 dans la construction de routes à travers le Canada. Mais en plus d'être très important du côté de l'ingénierie, il préoccupe tous ceux qui s'occupent de sécurité routière. Tout récent, ce problème est apparu à la fin des années 70. Il originerait des différentes modifications faites dans la construction des routes (ces changements ont été apportés dans le but de régler le problème de la fissuration des routes) et des changements observés au niveau de la circulation lors de ces années.

MAIS QU'EST-CE QUE L'ORNIÉRAGE[1]?

Pour bien comprendre ce qu'est l'orniérage, je vous conseille de jeter un coup d'oeil à la figure 1. Si on se fit à celle-ci, on peut définir l'orniérage comme étant une dépression longitudinale que l'on retrouve dans le sentier préférentiel des pneus des véhicules. De plus différentes études ont démontrés que ces dépressions pouvaient causer certains problèmes. Pour les différentes manoeuvre qu'un conducteur doit effectuer dans le cadre d'une conduite normale, ils vont de l'inconfort à l'accident pur et simple.

Figure 1 - Pictogramme d'ornières



(1): Un texte complémentaire (plus explicite) pour différentes parties du texte est disponible dans: Recuil des textes des présentations du colloque sur les méthodes et domaines d'application de la statistique, pp 319-328, MONTÉGIANI M.

Jusqu'ici la recherche a réussi à identifier trois types différents d'orniérage possible. Mais jusqu'à présent, aucune de celles-ci n'a clairement démontré les causes de l'orniérage. Ainsi, la recherche des causes devient une priorité.

RECHERCHE DES CAUSES

Les différents spécialistes du domaine ont évalués que les ornières peuvent être dues aux différents facteurs se rapportant à la construction d'une route ainsi qu'à son utilisation. Puisqu'il serait très difficile par une seule expérience de voir si tous ces éléments sont des causes d'ornières, un comité sur l'orniérage a été formé. Celui-ci a déterminé les facteurs les plus interressants à examiner et parmi ceux-ci les trois suivants ont étés retenus:

* le type de mélange;
* le volume de trafic;
* le nombre d'années depuis la construction de la route.

Chacun de ces facteurs selon le comité se compose de trois niveaux (cf tableau 1). La structure de chacun d'eux (facteurs croisés) permettait donc de penser à une planification d'expérience relativement courante.

Tableau 1 - Facteurs étudiés avec leurs différentes modalités

Facteurs	Modalités
Mélange	MB4 MB5 MB4R
Trafic	(Débit journalier moyen annuel) DJMA ≤ 10 000 10 000 < DJMA ≤ 20 000 20 000 < DJMA
Année	Moins de 3 ans Entre 3 et 6 ans Plus de 6 ans

PLANIFICATION DE L'EXPÉRIENCE

Pour savoir si chacun de ces facteurs avait une influence il suffisait d'utiliser un plan factoriel à trois facteurs, où les effets sont fixes. Ainsi, après avoir réparti les sections selon chacune des cellules (27) créées par le croisement de leurs modalités, il ne restait plus qu'à choisir de façon aléatoire les sections à étudier.

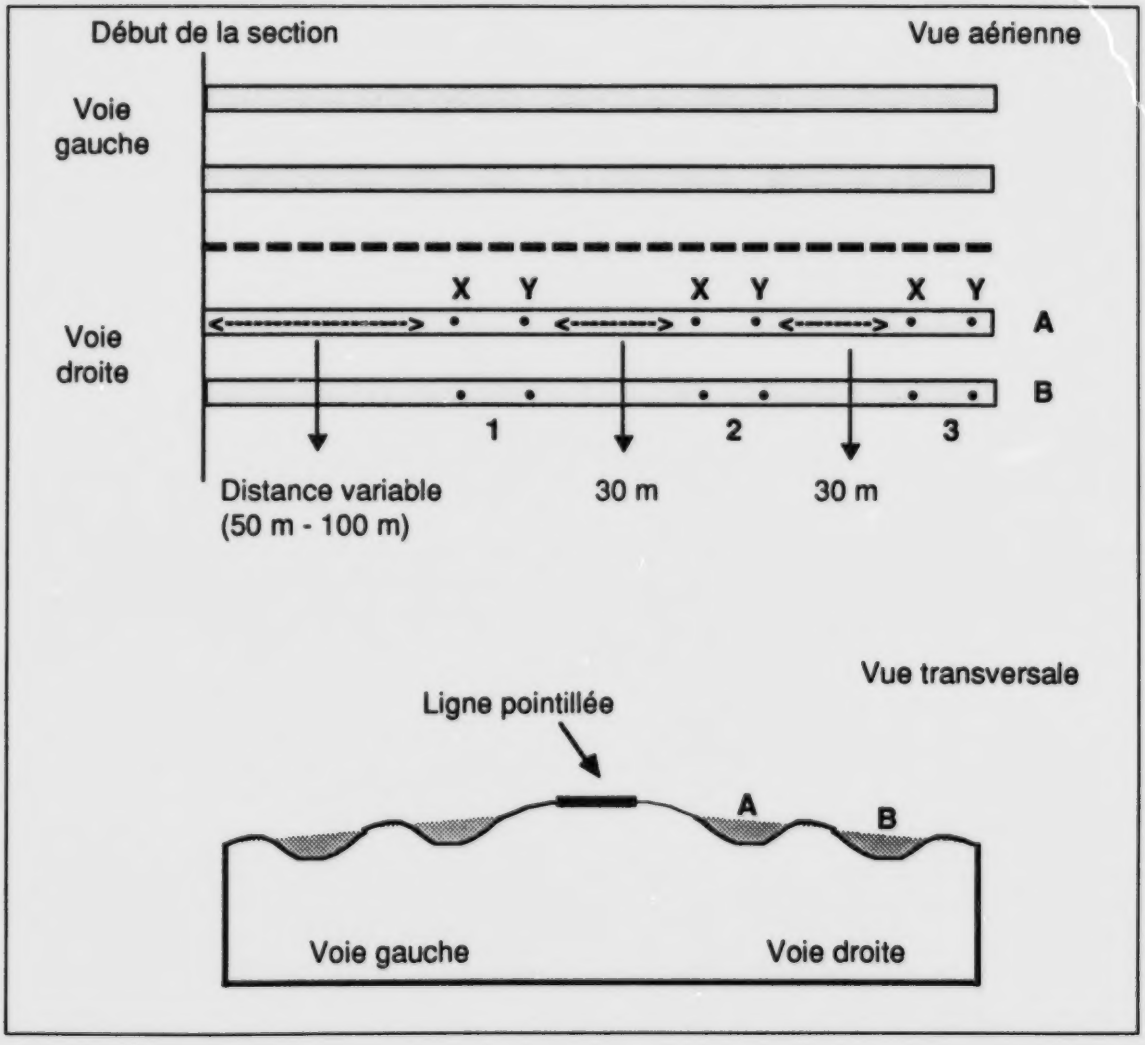
Le hasard fait bien souvent mal les choses, car lorsque les sections furent classées selon les cellules possibles certaines d'entre elles étaient vides ou ne contenaient qu'une section. Par contre, à ce moment, rien ne nous empêchait d'aller de l'avant avec notre projet car en se servant de la théorie des contrastes il était possible d'établir plusieurs comparaisons parmi les plus intéressantes. Ainsi, selon la disponibilité des sections nous avions pour l'expérience le schéma suivant (figure 2) où le nombre à l'intérieur de chacune des cellules correspond au nombre de sections choisies pour cette cellule .

Figure 2 - Schéma de l'expérience telle que planifiée

Mélange	Trafic	Année			Total
		Moins de 3 ans	Entre 3 et 6 ans	Plus de 6 ans	
MB4	JMA ≤10 000	2	1	0	3
	10 000< JMA ≤20 000	1	2	0	3
	20 000 < JMA	1	1	1	3
MB5	JMA ≤10 000	2	0	4	6
	10 000< JMA ≤20 000	1	1	1	3
	20 000 < JMA	4	1	0	5
MB4R	JMA ≤10 000	1	2	1	4
	10 000< JMA ≤20 000	1	1	0	2
	20 000 < JMA	1	0	1	2
Total		14	9	8	31

Cette planification ainsi faite il ne restait plus qu'à s'entendre sur les différents paramètres concercnant les unités expérimentales, c'est-à-dire: le choix, le mode de collecte et le nombre à recueillir. Après quelques rencontres il fût convenu que 12 carottes (prélèvements cylindriques allant de la surface vers le fond, d'une longeur de 30 cm et d'un diamètre de 10 cm), serait prises par section. Quant au mode de collecte, une distance variant entre 50 et 100 m à partir du début de la section et du premier point de cueillette, sera choisie au hasard. On prélèvera à ce point 4 carottes, deux dans chaque ornière (droite et gauche) de la voie de droite. Ensuite, 30 m plus loin il y aura de nouveau un point de cueillette où on en prendra 4 de la même façon qu'au premier point. Finalement 30 m plus loin, au dernier point, on en recueillera 4 autres (c.f. figure 3).

Figure 3 - Illustration du mode de collecte planifié



Donc tous les participants à cette expérience avait décidé de s'en tenir à cette planification qui, semblait pleine de promesse. Mais, bien souvent, dans le domaine expérimental, une planification aussi rigoureuse soit-elle, est confrontée à certains problèmes qui dépassent l'imaginable.

PROBLÈMES

Différents problèmes se sont abattus sur l'expérience. Les conséquences de ces problèmes sont plus ou moins fâcheuses et dans quelques cas ont empêché de réaliser l'expérience tel que prévue.

Choix des sections

Le choix des sections s'est fait à partir de l'inventaire fait en 1986 et les carottes ont été recueillies en 1987. Ce décalage a entraîné que plusieurs des sections échantillonnées ont étés modifiées. Celles-ci ont eu deux effets: soit faire changer de cellules certaines sections ou soit de les rendre inutilisables. Lorsqu'elles étaient inutilisables on essayait de trouver une section équivalente pour la remplacer, ce qui n'était pas toujours possible.

D'une façon plus générale l'impact de ces modifications fût d'augmenter le nombre de cellules vides. Ainsi au lieu d'avoir un schéma comparable à celui qu'on retrouvait à la figure 2 on a plutôt celui qu'on retrouve à la figure 4.

Figure 4 - Schéma de l'expérience telle que réalisée

Mélange	Trafic	Année			Total
		Moins de 3 ans	Entre 3 et 6 ans	Plus de 6 ans	
MB4	JMA ≤10 000	2	0	1	3
	10 000<JMA≤20 000	0	1	2	3
	JMA >20 000	0	0	2	2
MB5	JMA ≤10 000	2	2	2	6
	10 000< JMA ≤20 000	4	0	1	5
	20 000 < JMA	0	2	2	4
MB4R	JMA ≤10 000	1	0	1	2
	10 000< JMA ≤20 000	0	1	0	1
	20 000 < JMA	1	0	4	5
Total		10	6	15	31

Nombre d'unités expérimentales

Le temps est un déterminant dans une expérience. Puisque le domaine de l'étude se trouvait à l'extérieur, c'est-à-dire sur la route, et que les unités expérimentales devaient être recueillies sur celle-ci, il devenait important de faire cette cueillette dans un temps propice. Un temps propice, dans notre cas signifie de faire la cueillette le primptemps ou l'automne.

Puisque nous étions à la fin de l'automne, et que pour certaines raisons on ne pouvait pas attendre le primptemps, il fut décider par les gens qui faisaient la cueillette, pour des raisons de faisabilités, de ne cueillir que 6 unités expérimentales par section. Cette décision entraînait donc une diminution importante du nombre d'unité expérimentale. Ce qui, sans être dramatique, diminuait notre taille d'échantillon.

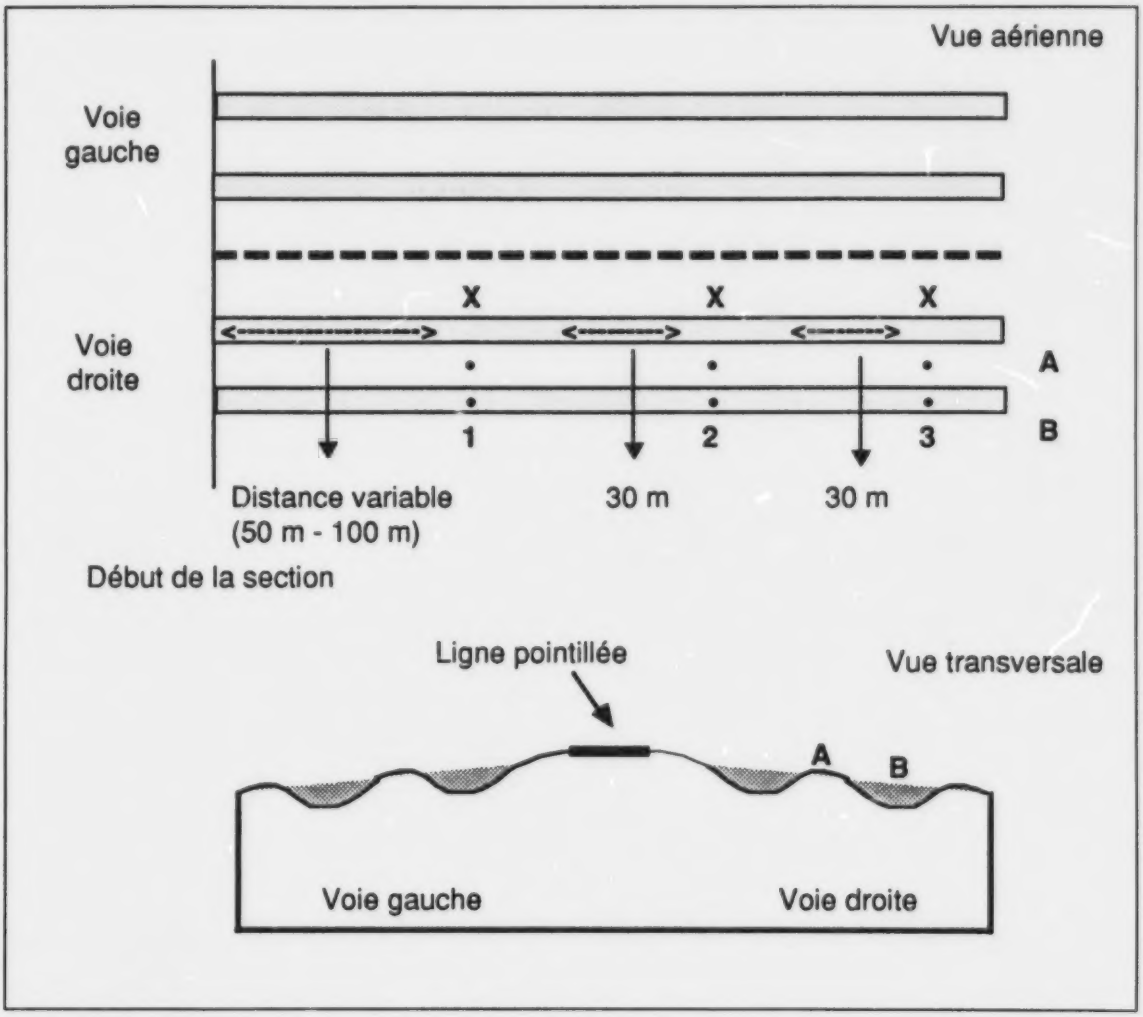
Mode de collecte

Bien que la façon de cueillir les carottes fût clairement établie dès la planification de l'enquête, un problème a quand même été décelé à cet endroit. Toutes les unités expérimentales devaient être prises dans le fond d'une ornière de façon à ce qu'une valeur de profondeur d'ornière soit associée à chacune des caractéristiques des unités expérimentales. Mais, au lieu de faire comme la planification le prévoyait, les gens qui recueillaient ces unités en ont pris une dans l'ornière gauche et ont pris l'autre dans le centre entre les deux ornières. Cette erreur, qui est dû à une erreur d'interprétation de l'illustration du mode de collecte, a changé bien des choses.

Ainsi, le respect du mode de collecte permettait de savoir si pour la voie de droite l'ornière de droite était plus importante que celle de gauche. Car selon une nouvelle théorie, ce phénomène pourait s'observer sur les autoroutes puisque les forces tangentielles imprimer par les véhicules éprouvent moins de résistance sur l'extérieur de la route que sur l'intérieur. Donc en suivant cette pensée, l'ornière droite de la voie droite serait plus importante que l'ornière gauche de la voie droite. Mais cette erreur devait enlever toutes possibilités de pousser plus loin nos investigations sur cette nouvelle théorie très contestée.

Donc au lieu de réaliser ce qui a été planifié, comme à la figure 3, l'éxécution a donné quelque chose de bien différent comme à la figure 5.

Figure 5 - Illustration du mode de collecte tel que réalisé



RÉSULTATS

Toutes ces catastrophes ont considérablement ralenti le processus de l'analyse qui devait débuter l'an dernier. À vrai dire, les seuls résultats qui sont disponibles sont les résultats principaux qui sont à l'origine de l'enquête.

Si nous gardons à l'esprit que le principal but poursuivi par cette expérience est de déterminer si les facteurs retenus ont une influence quelconque sur l'orniérage, alors le tableau 2 apporte les réponses.

Tableau 2 - Table d'analyse de la variance

Source	d.l.	Somme des Carrés	Moyenne des Carrés	Valeur de F	Pr >F
Modèle	16	35.80159937	2.23759996	6.71	0.0001
Erreur	76	25.32814519	0.33326507		
Total corrigé	92	61.12974456			

R-Carré	C.V.	Racine MCE	Moyenne des ornières
0.585666	26.18964	0.577291	2.20427311

Source	d.l.	Type I SC	Moyenne des Carrés	Valeur de F	Pr > F
MÉLANGE	2	2.39583265	1.19791633	3.59	0.0322
TRAFFIC	2	1.12741220	0.56370610	1.69	0.1911
MÉLANGE*TRAFFIC	4	9.81983617	2.45495904	7.37	0.0001
ANNÉE	2	1.48041674	0.74020837	2.22	0.1155
MÉLANGE*ANNÉE	3	14.40390995	4.80130332	14.41	0.0001
TRAFFIC*ANNÉE	3	6.57419165	2.19139722	6.58	0.0005
MÉLANG*TRAFFIC*ANNÉE	0	0.00000000	.	.	.

Source	d.l.	Type IV SC	Moyenne des Carrés	Valeur de F	Pr > F
MÉLANGE	2*	4.54330207	2.27165103	6.82	0.0019
TRAFFIC	2*	8.12102171	4.06051086	12.18	0.0001
MÉLANGE*TRAFFIC	3	9.04182354	3.01394118	9.04	0.0001
ANNÉE	2*	1.17784613	0.58892307	1.77	0.1778
MÉLANGE*ANNÉE	2	4.83477714	2.41738857	7.25	0.0013
TRAFFIC*ANNÉE	3	6.57419165	2.19139722	6.58	0.0005
MÉLANG*TRAFFIC*ANNÉE	0	0.00000000	.	.	.

* NOTE: D'autres hypothèses testables de type IV existent lorsque produit avec différentes SC.

Facteurs

Le tableau 2 indique que les facteurs comme le trafic et le mélange influencent directement l'orniérage, alors que le facteur année, pour sa part, n'influence pas l'orniérage.

Trafic : Considéré par plusieurs spécialistes comme le seul facteur ayant de l'importance dans la formation d'ornière le résultat appraît comme normal.

Mélange: Ce résultat confirme la supposition qui prétend que le matériau utilisé est de la plus haute importance quand il s'agit de la construction d'une route.

Année : En fait, bien qu'intuitivement il semble légitime de croire que ce facteur n'a pas d'influence, il semblait en être tout autrement pour certains spécialistes. Malgré tout, l'analyse prouve que ce facteur n'a pas l'importance que ceux-ci lui accordait.

Intéractions

Dans une expérience de ce genre l'intéraction possible entre deux facteurs est toujours intéressante à regarder et à analyser. Ce cas-ci ne fait pas exception à la règle, certains résultats surprennent.

Trafic*Mélange : Le fait que cette intéraction soit significative semble provenir du faitque certains matériaux sont mieux adaptés pour certaines conditions de circulation. Ce qui correspond à la théorie qui considère que pour les trafic les plus denses, il est préférable d'utiliser certains matériaux.

Trafic*Année : C'est une surprise de voir que cette intéraction est significative. Le fait qu'elle soit significative peut s'expliquer par l'inadéquation de la variable suggérée pour mesurer l'intensité du trafic. En prenant le trafic cumulé depuis le début de la construction de la route plutôt que le trafic jounalier moyen la mesure aurait été bien plus juste.

Mélange*Année : Il est surprenant de constater que cette intéraction est significative. Par contre, ce résultat confirme que certains mélange s'oxydent et se déteriorent plus rapidement que d'autres (surtout le cas des mélanges recyclés MB4R).

Trafic*Année*Mélange: Étant donné le genre de l'expérience cette intéraction n'est pas estimable. Il est donc difficile de vouloir en tirer des conclusions.

CONCLUSION

A la lumière des résultats précédents, il semble que la détermination des causes d'orniérages ne soit pas réellement terminée. En effet, bien que certaines causes soient identifiées il n'en demeure pas moins quelques interrogations qui laissent envisager d'autres collectes de données sur le sujet.

Donc, bien qu'il soit probable, qu'encourager par les résultats obtenus les demandeurs fassent une autre expérience de ce genre la fin de ce projet n'est pas imminente. Par contre dans le cadre du colloque sur les méthodes et domaines d'application de la statistique cette communication boucle la boucle à propos de la production d'information statistique grâce à une planification d'expérience.

Bibliographie:

FREUND R.J. et LITELL R. (1981), Sas for linear models: A guide to the ANOVA and GLM procedures, Sas Institute Inc., North Carolina, ISBN 0-917382-31-5

KISH L. (1987), Statistical design for Research, John Wiley and sons, Toronto, ISBN 0-471-08359-3

MILIKEN G.A. ET JOHNSON D.E.(1984), Analysis of Messy Data volume 1: Designed Experments, Van Nostrand Reinhold Company, New York (N.Y.), ISBN 0-534-02713-7

MONTGOMERY D.C. (1984), Design and analysis of experiments, seconde édition, John Wiley and sons, Toronto, ISBN 0-471-86812-4

NETER J. et WASSERMAN W. (1974), Applied linear statistical models, Richard D. Irwin, Homewood (Illinois), ISBN 0-256-01498-1

PURI M.L., VILAPLANA J.P. et WERTZ W. (1987), New perspectives in theorical and Applied Statistics, John Wiley and sons, Toronto, ISBN 0-471-84800-X

SESSION SUR LA BIOSTATISTIQUE

Estimation simultanée de la fiabilité

intra-juges et inter-juges

Gina Bravo
Centre de recherche en gériatrie
Hôpital d'Youville
Sherbrooke (Québec)
J1H 4C4

En recherche clinique, il est souvent nécessaire de développer un instrument de mesure adapté spécifiquement à la condition qu'on cherche à mesurer. Tout développement d'un nouvel instrument doit être suivi d'une étude de sa validité et de sa fiabilité.

Le coefficient de corrélation intra-classe a été proposé par Fisher pour estimer les fiabilités intra- et inter-juges. Celles-ci sont estimées à partir d'une analyse de la variance (ANOVA) à deux facteurs pour mesures répétées.

Nous proposons, pour le cas d'un modèle aléatoire, une façon originale d'estimer simultanément ces deux fiabilités à l'aide d'une ANOVA à trois facteurs. En nous inspirant de la méthode de Satterthwaite pour approximer la distribution d'une combinaison linéaire de Chi^2, nous proposons également un intervalle de confiance autour du paramètre estimé.

MOTS-CLES: Coefficient intra-classe; reproductibilité; intervalle de confiance.

1. INTRODUCTION

Tout développement d'un nouvel instrument de mesure conçu pour évaluer l'efficacité d'un traitement doit être suivi de l'évaluation des qualités psychométriques de l'instrument (Carmines et Zeller 1979). Deux qualités essentielles que doit posséder tout instrument de mesure sont la validité et la reproductibilité (appelée aussi fiabilité ou fidélité). Vérifier la validité d'un instrument consiste à s'assurer qu'il mesure bien ce qu'il doit mesurer. La reproductibilité d'autre part, est la capacité de l'instrument à fournir des résultats équivalents lorsqu'il est utilisé à plusieurs reprises sur un même sujet dans des conditions constantes. La théorie classique décrit deux types de reproductibilité: inter-juges et intra-juges. Comme son nom

l'indique, la fiabilité inter-juges s'évalue en comparant les résultats obtenus par plusieurs individus utilisant l'instrument sur un même sujet. D'autre part, la fiabilité intra-juges concerne la comparaison des résultats obtenus par un seul individu utilisant l'instrument sur un même sujet à deux moments différents, l'intervalle de temps séparant les deux mesures étant suffisamment court pour permettre de supposer que la condition mesurée n'a pas changé.

2. DEFINITION DE LA REPRODUCTIBILITE

Soit X la valeur observée, V la vraie valeur et e la différence entre l'observation et la réalité. La définition de la reproductibilité est basée sur le modèle linéaire simple: $X = V + e$. Les données d'une étude de fiabilité peuvent être représentées sous la forme suivante:

		Mesure			
		M_1	M_2	...	M_k
Sujet	S_1	X_{11}	X_{12}	...	X_{1k}
	S_2	X_{21}	X_{22}	...	X_{2k}
	.	.	.		.
	.	.	.		.
	S_n	X_{n1}	X_{n2}	...	X_{nk}

où la variable X_{ij} représente la j^e observation sur le i^e sujet et est modélisée par l'équation $X_{ij} = V_i + e_{ij}$. Les hypothèses que doivent satisfaire les variables aléatoires du modèle sont celles sur lesquelles s'appuient l'analyse de la variance, à savoir:

a) $E(e) = 0$ et $Var(e) = \sigma^2_e$,
b) $E(V) = \mu$ et $Var(V) = \sigma^2_V$,
c) l'indépendance entre V et e.

L'hypothèse d'indépendance entre la valeur à mesurer et l'erreur aléatoire entraîne la décomposition suivante de la variation observée:

$\sigma^2_X = \sigma^2_V + \sigma^2_e$. Le résultat obtenu à l'aide d'un instrument sera reproductible dans la mesure où σ^2_e sera petit relativement à σ^2_V; d'où la définition de la fiabilité:

$$R = \frac{\sigma^2_V}{\sigma^2_X} = \frac{\sigma^2_V}{\sigma^2_V + \sigma^2_e}.$$

Clairement, R varie entre 0 et 1. La mesure est reproductible lorsque $\sigma^2_e = 0$, donc lorsque R = 1. A l'opposé, le cas extrême d'une variation infinie de l'erreur aléatoire entraîne une fiabilité nulle. L'idéal consiste donc à viser une fiabilité de 1. Cet idéal n'étant jamais atteint en pratique, il est d'usage de considérer comme acceptable une fiabilité de l'ordre de 90%.

3. ESTIMATION DE LA REPRODUCTIBILITE

L'estimation de la reproductibilité consiste à estimer les termes apparaissant dans sa définition, soit σ^2_V et σ^2_e. L'outil statistique utilisé est l'analyse de la variance pour mesures répétées. Lorsque n sujets choisis au hasard sont mesurés par k juges différents, une analyse de la variance à deux facteurs fournit l'information décrite à la table suivante:

Table 1. Analyse de la variance à deux facteurs

Source de variation	SC	dl	MC	E(MC)*
Inter-sujets	$k\Sigma(\bar{X}_{i.}-\bar{X}_{..})^2$	n-1	MC_S	$\sigma^2_e+k\sigma^2_V$
Intra-sujets				
Juge	$n\Sigma(\bar{X}_{.j}-\bar{X}_{..})^2$	k-1	MC_J	$\sigma^2_e+n\sigma^2_J$
Erreur	par soustraction	(n-1)(k-1)	MC_e	σ^2_e
Totale	$\Sigma(X_{ij}-\bar{X}_{..})^2$	nk-1		

*Espérances des moindres carrés pour un modèle aléatoire.

Le modèle général sur lequel est basé cette analyse est défini par $X_{ij} = V_i + J_j + e_{ij}$ où J_j représente l'effet additif du j^e juge. Les hypothèses habituelles d'indépendance entre les différentes variables entraînent que $\sigma^2_X = \sigma^2_V + \sigma^2_J + \sigma^2_e$. Il s'en suit alors que:

$$R = \frac{\sigma^2_V}{\sigma^2_X} = \frac{\sigma^2_V}{\sigma^2_V + \sigma^2_J + \sigma^2_e}. \qquad (3.1)$$

L'estimation des termes σ^2 se fait à partir des espérances des moindres carrés apparaissant à l'extrême droite de la table précédente:

$$E(MC_e) = \sigma^2_e \quad \Rightarrow \quad \hat{\sigma}^2_e = MC_e,$$
$$E(MC_J) = \sigma^2_e + n\sigma^2_J \quad \Rightarrow \quad \hat{\sigma}^2_J = (MC_J - MC_e) / n,$$
$$E(MC_S) = \sigma^2_e + k\sigma^2_V \quad \Rightarrow \quad \hat{\sigma}^2_V = (MC_S - MC_e) / k.$$

Après substitution de ces estimateurs dans l'équation (3.1), nous obtenons:

$$\hat{R} = \frac{\hat{\sigma}^2_V}{\hat{\sigma}^2_V + \hat{\sigma}^2_J + \hat{\sigma}^2_e} = \frac{n\,(MC_S - MC_e)}{n\,MC_S + k\,MC_J + (nk-n-k)\,MC_e}.$$

L'estimateur $\hat{R}$ de la fiabilité est appelé le coefficient de corrélation intra-classe et a d'abord été proposé par Bartko (1966).

3.1 Exemple

Les données figurant aux tables 2 et 3 de la page suivante résument six observations effectuées sur vingt sujets. Il s'agit de mesures de force enregistrées à l'aide d'un myomètre au cours de l'étirement passif des muscles fléchisseurs de la cheville et du coude. En comparant les variances des deux ensembles de données, on constate que celles-ci sont beaucoup plus élevées dans le premier cas que dans le deuxième. Cette impression ressort également de l'observation des figures 1 et 2 où les six mesures par sujets sont représentées graphiquement. Tout semble donc indiquer que la mesure prise sur la cheville est moins reproductible que celle effectuée sur le coude. En soumettant les données à une analyse de la variance pour mesures répétées, cette impression est confirmée puisque $\hat{R}$ = 61% pour la cheville et $\hat{R}$ = 86% pour le coude.

Table 2. Mesures de la force musculaire de la cheville

Sujet	Moyenne*	Variance
1	9.707	0.6623
2	6.077	1.5229
3	5.985	0.7292
4	8.415	4.1603
5	5.227	1.4739
6	12.555	6.9022
7	8.368	5.5309
8	10.438	14.1960
9	8.265	3.9071
10	12.617	12.6792
11	13.410	5.4662
12	15.247	7.5628
13	13.338	8.0554
14	13.783	9.5936
15	14.062	9.2212
16	10.085	3.4142
17	12.920	0.5310
18	12.685	2.2451
19	10.500	0.8827
20	9.290	0.6738

*Moyenne de six observations par sujet
Source: Malouin et al. (1989)

Table 3. Mesures de la force musculaire du coude

Sujet	Moyenne	Variance
1	1.090	0.0373
2	1.812	0.2096
3	3.103	1.1769
4	0.625	0.0318
5	1.765	0.1030
6	0.697	0.0348
7	6.100	0.4960
8	2.323	0.2037
9	1.682	0.1475
10	3.738	0.6356
11	3.518	0.5462
12	1.923	0.3650
13	0.230	0.0096
14	0.977	0.1401
15	1.527	0.9441
16	1.738	0.3602
17	1.418	0.1084
18	0.217	0.0087
19	2.357	0.1269
20	1.498	0.0173

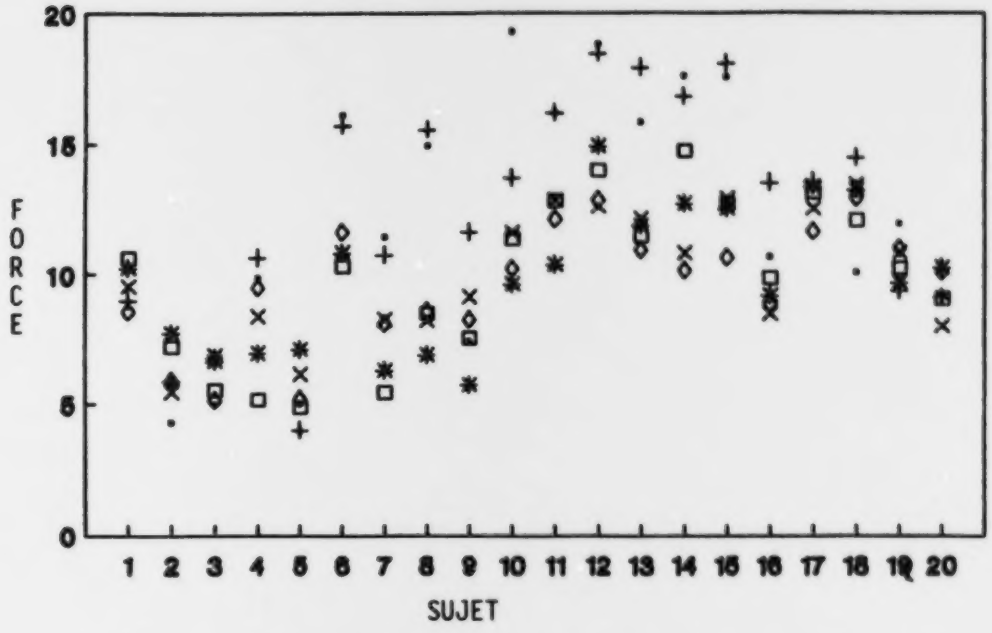


Figure 1. Mesures de la force musculaire de la cheville. Les differents symboles représentent les six mesures prises sur chaque sujet.

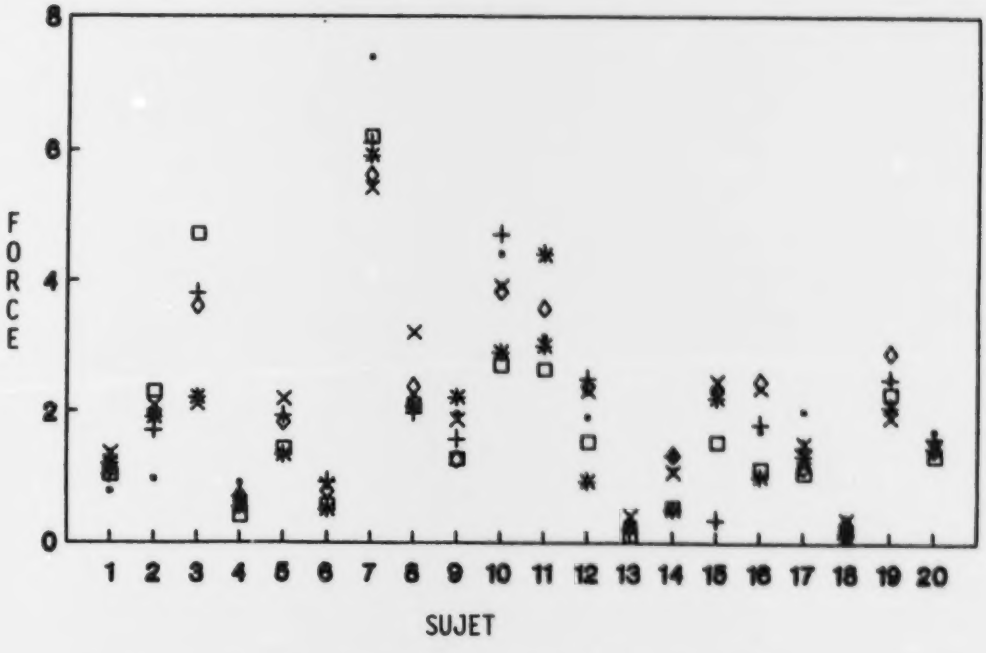


Figure 2. Mesures de la force musculaire du coude.

3.2 Intervalle de confiance

Pour tenir compte de la variation échantillonnale, un intervalle de confiance devrait être associé à tout estimateur ponctuel de la fiabilité. Un intervalle approximatif a été proposé par Fleiss et Shrout (1978). La version unilatérale de celui-ci est défini par

$$R \geq \frac{n\,(MC_S - F_{\alpha;n-1,v}\,MC_e)}{n\,MC_S + F_{\alpha;n-1,v}\,[k\,MC_J + (nk-n-k)MC_e]}$$

avec $F = MC_J / MC_e$ et $v = \dfrac{(n-1)(k-1)\,[k\,\hat{R}\,F + n(1+(k-1)\hat{R} - k\hat{R}]^2}{(n-1)k^2\,\hat{R}^2\,F^2 + [n(1+(k-1)\hat{R}) - k\hat{R}]^2}$.

En utilisant ces expressions pour obtenir un intervalle de confiance unilatéral au niveau 95% de la fiabilité de la mesure prise sur le coude, nous obtenons $R \geq 79\%$.

4. MODELE A TROIS FACTEURS

La façon dont les deux exemples précédents ont été présentés laisse croire que les vingt sujets ont été mesurés par six individus différents. Il s'agit en fait de trois techniciens qui ont mesuré tous les sujets à deux reprises. La vraie structure des données est mise en évidence par la table 4 de la page suivante. Cette base de données permet d'estimer plusieurs indices de reproductibilité: deux fiabilités inter-juges (une au temps 1 et une seconde au temps 2), et trois fiabilités intra-juges (une associée à chacun des juges). Donc au total, cinq indices de fiabilité seront obtenus. Or, on peut souhaiter vouloir combiner ces valeurs en un indice global et unique de reproductibilité.

Pour estimer simultanément la reproductibilité inter- et intra-juges, il suffit d'utiliser l'analyse de la variance à trois facteurs, plutôt que plusieurs analyses de la variance à deux facteurs. Il est clair en effet, que le devis sous-jacent aux données de la table 4 est un plan d'expérience à trois facteurs et que la variation des mesures d'un sujet devrait être expliquée par le facteur "juge" et le facteur

"temps", le reste étant inexpliqué et dû à l'erreur aléatoire. De plus, il est souhaitable de tenir compte des effets d'interaction entre les différents facteurs.

Table 4. Structure réelle des données recueillies pour mesurer la force des muscles fléchisseurs du coude

	Temps 1			Temps 2		
Sujet	Juge 1	Juge 2	Juge 3	Juge 1	Juge 2	Juge 3
1	0.78	1.20	1.35	1.03	1.03	1.15
2	0.97	1.90	2.07	1.70	2.30	1.93
3	2.22	2.20	2.10	3.80	4.70	3.60
4	0.90	0.55	0.50	0.70	0.40	0.70
5	1.90	1.33	2.17	1.93	1.43	1.83
6	0.53	0.50	0.87	0.93	0.57	0.78
7	7.40	5.90	5.40	6.10	6.20	5.60
8	2.23	2.07	3.20	1.97	2.10	2.37
9	1.95	2.20	1.85	1.57	1.27	1.25
10	4.40	2.90	3.90	4.70	2.70	3.83
11	3.13	3.00	4.38	4.40	2.63	3.57
12	1.90	0.93	2.30	2.50	1.53	2.38
13	0.23	0.20	0.40	0.20	0.10	0.25
14	1.10	0.50	1.07	1.33	0.53	1.33
15	0.33	2.20	2.45	0.35	1.53	2.30
16	1.70	1.00	2.35	1.80	1.13	2.45
17	2.00	1.33	1.50	1.45	1.07	1.16
18	0.20	0.10	0.35	0.15	0.20	0.30
19	2.50	2.07	1.90	2.50	2.27	2.90
20	1.70	1.45	1.48	1.60	1.33	1.43

Une analyse de la variance à trois facteurs décompose la variation totale de la façon suivante:

Table 5. Analyse de la variance à 3 facteurs*

Source de variation	dl	MC	E(MC)
Inter-sujets	n-1	MC_S	$\sigma^2_e+t\sigma^2_{SJ}+k\sigma^2_{ST}+tk\sigma^2_V$
Intra-sujets			
Juge	k-1	MC_J	$\sigma^2_e+t\sigma^2_{SJ}+n\sigma^2_{JT}+tn\sigma^2_J$
Temps	t-1	MC_T	$\sigma^2_e+k\sigma^2_{ST}+n\sigma^2_{JT}+kn\sigma^2_T$
Sujet x Juge	(n-1)(k-1)	MC_{SJ}	$\sigma^2_e + t\sigma^2_{SJ}$
Sujet x Temps	(n-1)(t-1)	MS_{ST}	$\sigma^2_e + k\sigma^2_{ST}$
Juge x Temps	(k-1)(t-1)	MS_{JT}	$\sigma^2_e + n\sigma^2_{JT}$
Erreur	(n-1)(k-1)(t-1)	MS_e	σ^2_e

*Pour simplifier la présentation, la colonne des SC n'est pas reportée.

L'équation générale décrivant une observation X_{ijl} prise par le juge j au temps l sur le sujet i, est donnée par $X_{ijl} = V_i + J_j + T_l + (SJ)_{ij} + (ST)_{il} + (JT)_{jl} + e_{ijl}$ et la variation observée est décrite par l'expression $\sigma^2_X = \sigma^2_V + \sigma^2_J + \sigma^2_T + \sigma^2_{SJ} + \sigma^2_{ST} + \sigma^2_{JT} + \sigma^2_e$. La fiabilité est alors définie par

$$R = \frac{\sigma^2_V}{\sigma^2_V + \sigma^2_J + \sigma^2_T + \sigma^2_{SJ} + \sigma^2_{ST} + \sigma^2_{JT} + \sigma^2_e} . \qquad (4.1)$$

Comme pour le cas bi-dimensionnel, les espérances des moindres carrés servent à estimer les différentes composantes de la variance:

$$\begin{aligned}
\hat{\sigma}^2_e &= MC_e, \\
\hat{\sigma}^2_{JT} &= (MC_{JT} - MC_e) / n, \\
\hat{\sigma}^2_{ST} &= (MC_{ST} - MC_e) / k, \\
\hat{\sigma}^2_{SJ} &= (MC_{SJ} - MC_e) / t, \\
\hat{\sigma}^2_T &= (MC_T - MC_{ST} - MC_{JT} + MC_e) / kn, \\
\hat{\sigma}^2_J &= (MC_J - MC_{SJ} - MC_{JT} + MC_e) / nt, \\
\hat{\sigma}^2_V &= (MC_S - MC_{SJ} - MC_{ST} + MC_e) / kt,
\end{aligned}$$

ce qui, après simplification, conduit à l'estimateur suivant:

$$\hat{R} = \frac{n(MC_S - MC_{SJ} - MC_{ST} + MC_e)}{\begin{aligned}(n\,MC_S + k\,MC_J + t\,MC_T + (nk-n-k)MC_{SJ} + (tn-t-n)MC_{ST} + \\ (tk-t-k)MC_{JT} + [(nk-n-k)(t-1) + t]MC_e).\end{aligned}} \qquad (4.2)$$

4.1 Exemple

Une analyse de la variance à trois facteurs effectuée sur les données de la table 5 génère l'information suivante:

Table 6. Analyse de la variance à trois facteurs associée aux données de la table 5

Source de variation	SC	dl	MC
Inter-sujets	217.84	19	11.465
Intra-sujets			
Juge	2.67	2	1.335
Temps	0.11	1	0.110
Sujet x Juge	15.00	38	0.395
Sujet x Temps	7.11	19	0.374
Juge x Temps	0.12	2	0.060
Erreur	3.50	38	0.092
Totale	246.35	119	

En remplaçant les moindres carrés de l'équation (4.2) par leur valeur numérique, la fiabilité globale est estimée à 83%.

4.2 Intervalle de confiance

Lorsqu'une fiabilité acceptable est obtenue, on peut souhaiter tester une hypothèse concernant la valeur réelle de la fiabilité ou construire un intervalle de confiance autour de la valeur obtenue.

L'hypothèse d'une fiabilité nulle étant équivalente à l'hypothèse H_0: $\sigma^2_V = 0$, le test s'obtient directement de la table de l'analyse de la variance:

$$H_0: R = 0 \text{ rejetée si } F = MC_S / MC_e > F_{\alpha;n-1,(n-1)(k-1)(t-1)}.$$

L'intervalle de confiance autour de R est cependant plus compliqué à obtenir puisqu'il est fonction non seulement de MC_S et MC_e, mais également des autres moindres carrés.

La description qui suit fourni les détails de la construction d'un intervalle de confiance approximatif autour de la vraie valeur de la fiabilité. La méthodologie étend celle proposée par Fleiss et Shrout (1978) pour le cas bi-dimensionnel.

Utilisant l'équation (4.1) qui défini la fiabilité, nous pouvons écrire

$$\sigma^2_V = \frac{R}{1-R}(\sigma^2_J + \sigma^2_T + \sigma^2_{SJ} + \sigma^2_{ST} + \sigma^2_{JT} + \sigma^2_e).$$

En substituant à σ^2_V son expression en terme de R, $E(MC_S)$ peut être exprimé sous la forme:

$$E(MC_S) = \frac{1}{1-R}\{\sigma^2_e(1+(tk-1)R) + \sigma^2_{SJ}t(1+(k-1)R) + \sigma^2_{ST}k(1+(t-1)R) + tkR(\sigma^2_J + \sigma^2_T + \sigma^2_{JT})\}.$$

Considérons maintenant la variable aléatoire suivante:

$$V = \frac{1}{1-R}\{\hat{\sigma}^2_e(1+(tk-1)R) + \hat{\sigma}^2_{SJ}t(1+(k-1)R) + \hat{\sigma}^2_{ST}k(1+(t-1)R) + tkR(\hat{\sigma}^2_J + \hat{\sigma}^2_T + \hat{\sigma}^2_{JT})\}.$$

Clairement $E(V) = E(MC_S)$. De plus, nous savons que $MC_S \sim Chi^2_{n-1}$ et que V est indépendante de MC_S. Par conséquent, s'il était possible de montrer que V suit également une distribution du Chi^2, avec un nombre v de degrés de liberté, nous aurions que $MC_S / V \sim F_{n-1,v}$ et de l'expression

$$1 - \alpha = Pr\,(MC_S / V < F_{\alpha;n-1,v}),$$

nous pourrions construire un intervalle de confiance pour R puisque V est fonction de R. En utilisant le théorème de Satterthwaite (1946), il est possible de montrer que V suit approximativement une distribution du Chi^2.

Théorème de Satterthwaite: Soient S^2_1, S^2_2, ..., S^2_n des variables aléatoires indépendantes telles que $S^2_i \sim Chi^2_{v_i}(\sigma^2_i / v_i)$. Alors $\Sigma a_i S^2_i \approx Chi^2_{v'}(a / v')$ où $a = \Sigma a_i \sigma^2_i$ et $v' = (\Sigma a_i \sigma^2_i)^2 / \Sigma[(a_i \sigma^2_i)^2 / v_i]$.

L'application de ce théorème permet d'affirmer que $MC_S / V \approx F_{n-1,v}$ où v peut être estimé par

$$\frac{[1/n(1-\hat{R})][a_1 MC_e + a_2 MC_{SJ} + a_3 MC_{ST} + a_4 MC_{JT} + a_5 MC_T + a_6 MC_J]^2}{\frac{(a_1 MC_e)^2}{(n-1)(k-1)(t-1)} + \frac{(a_2 MC_{SJ})^2}{(n-1)(k-1)} + \frac{(a_3 MC_{ST})^2}{(n-1)(t-1)} + \frac{(a_4 MC_{JT})^2}{(k-1)(t-1)} + \frac{(a_5 MC_T)^2}{(t-1)} + \frac{(a_6 MC_J)^2}{(k-1)}},$$

les constantes a_i étant définies par:

$$a_1 = \hat{R}(n-1)(tk-t-k) - n(\hat{R}+1),$$
$$a_2 = \hat{R}(nk-n-k) + n,$$
$$a_3 = \hat{R}(nt-n-t) + n,$$
$$a_4 = \hat{R}(tk-t-k),$$
$$a_5 = \hat{R}t,$$
$$a_6 = \hat{R}k.$$

De ce résultat, nous obtenons le corollaire suivant:

Corollaire: La borne inférieure de la fiabilité au niveau α peut être approximée par l'expression suivante:

$$R \geq \frac{n(MC_S - F^*(MC_{SJ} + MC_{ST} - MC_e)}{n\,MC_S + F^*\{[(n-1)(tk-t-k) - n]MC_e + (nk-n-k)MC_{SJ} + (nt-n-t)MC_{ST} + (tk-t-k)MC_{JT} + tMC_T + kMC_J\}}$$

où $F^* = F_{\alpha;n-1,v}$.

4.3 Exemple

L'application de ce corollaire aux données de la table 5 nous fournit une borne inférieure de 54% pour un seuil de confiance de 95%.

5. DISCUSSION

La faible valeur de la borne inférieure peut semer un doute quant à la valeur du corollaire. Des études de simulation faisant appel à la technique du bootstrap seront entreprises sous peu pour étudier le comportement de l'intervalle. Il est à noter cependant, que la validité de l'intervalle de confiance s'appuit sur la justesse de l'approximation de Satterthwaite et que celle-ci a été étudiée par des techniques de simulation (Gaylor et Hopper 1969) ainsi que par des analyses mathématiques exactes (Fleiss 1971).

BIBLIOGRAPHIE

Bartko, J.J. (1966), "The intraclass correlation coefficient as a measure of reliability," Psychological Reports, 19, 3-11.

Carmines, E.G. et Zeller, R.A. (1979), Reliability and validity assessment (Ser 17), Sage University Paper Series and Quantitative Application in the Social Sciences, Beverly Hills, CA: Sage.

Fleiss, J.L. (1971), "On the distribution of a linear combination of independent chi squares," Journal of the American Statistical Association, 66, 142-144.

Fleiss, J.L. et Shrout, P.E. (1978), "Approximate interval estimation for a certain intraclass correlation coefficient," Psychometrika, 43, 259-262.

Gaylor, D.W. et Hopper, F.N. (1969), "Estimating the degrees of freedom for linear combinations of mean squares by Satterthwaite's formula," Technometrics, 11, 691-705.

Malouin, F., Boiteau, M., Bonneau, C., Pichard, L. Bravo, G. (1989), "Use of a hand-held dynamoneter for the evaluation of spasticity in a clinical setting: a reliability study," Physiotherapy Canada, 41, 126-134.

Satterthwaite, F.E. (1946), "An approximate distribution of estimates of variance components," Biometrics, 2, 110-114.

ESTIMATION POUR PETITES REGIONS EN SANTE COMMUNAUTAIRE

Sylvie Cardin*, Samy Suissa* et Jean-François Boivin*

Résumé

Le Québec est divisé en 32 départements de santé communautaire couvrant une population de 6 000 000 d'habitants. Ces départements ont comme mandat, entre autres, d'établir le profil de santé de leur population et d'élaborer des programmes de prévention. Les données requises à ces fins peuvent être obtenues lors d'enquêtes locales ou à l'aide de banques de données nationales. Dans ce dernier cas, on doit avoir recours aux estimateurs pour petites régions. Plusieurs de ces estimateurs ont été proposés dans la littérature statistique, mais très peu d'entre eux ont été évalués empiriquement, particulièrement dans le domaine de la santé. Nous conduisons présentement une telle étude d'évaluation, dans laquelle des estimés pour de petites régions sont obtenus à partir d'une enquête nationale en santé et comparés à des estimés obtenus directement dans la petite région, lors de la répétition de l'enquête nationale à l'échelle locale. Les étapes de l'étude sont discutées.

* Département d'épidémiologie et de biostatistique de l'Université McGill et
Département de santé communautaire de l'Hôpital Général de Montréal.

Ce projet est subventionné conjointement par le Fonds de la Recherche en Santé du Québec (FRSQ) et le Programme National de Recherche et de Développement en Matière de Santé (PNRDS).

Ce travail concerne une étude comparative de l'exactitude empirique de différents estimateurs pour petites régions (EPR). Nous décrirons d'abord pourquoi ces estimateurs sont utilisés, les définirons et expliquerons, de façon générale, comment ils fonctionnent. Nous indiquerons ensuite pourquoi ce travail a été entrepris et en établirons la stratégie. Nous terminerons en donnant quelques exemples d'EPR.

1. Utilisation, définition et fonctionnement des estimateurs pour petites régions

Les départements de santé communautaire (DSC) ont souvent besoin de données concernant l'état de santé des membres de leur communauté. Quand cette situation se présente, trois solutions sont possible:

1- Les données du plus récent recensement peuvent être utilisées. Malheureusement, seules certaines variables socio-démographiques sont mesurées lors du recensement et leur pertinence dans le contexte d'une étude particulière peut laisser à désirer. De plus, quand l'étude doit être réalisée pour une année non-recensée, les utilisateurs font face au problème de l'extrapolation.

2- Une enquête locale peut être conduite. Les DSC n'ont cependant pas toujours les ressources nécessaires à la conduite de telles enquêtes dans leur communauté.

3- Les données provenant d'enquêtes nationales ou provinciales en santé peuvent être utilisées. Dans ce dernier cas, on doit avoir recours aux estimateurs pour petites régions.

Les estimateurs pour petites régions constituent donc un groupe de méthodes statistiques utilisées pour produire des estimés pour de petites régions en utilisant et combinant l'information provenant de différentes sources: vastes enquêtes (nationales, provinciales), registres et recensements.

En général, les EPR fonctionnent de la façon suivante. Supposons que nous voulions connaître la consommation moyenne d'alcool dans une petite région, Montréal par exemple. Nous connaissons déjà, au niveau national, la consommation moyenne d'alcool par groupe d'âge-sexe-statut socio-économique(sse) (nous avons obtenu ces données lors d'une enquête nationale en santé). En combinant cette information à la distribution par groupes d'âge-sexe-statut socio-économique de la petite région d'intérêt (ces données provenant d'un recensement ou d'une enquête locale réalisée antérieurement), nous obtenons la consommation moyenne d'alcool ESPEREE dans la petite région. La qualité de cette estimation dépend entièrement de la pertinence de l'utilisation des données nationales pour la production de l'estimé local.

2. Pourquoi une étude d'évaluation des estimateurs pour petites régions en santé?

Depuis que les démographes se sont penché sur la question des EPR, plusieurs méthodes ont été mises au point: estimateur simple direct, synthétique, de régression, composé, SPREE, etc. Ces estimateurs diffèrent, entres autres, selon l'approche prise pour les développer (approche prédictive ou approche classique dépendante de l'échantillonnage), la distribution sous-jacente assumée (Gaussienne, Poisson ou autre), les caractéristiques des paramètres que l'on estime (fixes ou ayant eux-même une distribution) et les critères à optimiser (moindres carrés, maximum de vraisemblance, etc.). Certains de ces EPR sont fréquemment utilisés bien que très peu[1] eussent été évalués empiriquement, particulièrement dans le domaine de la santé, i.e., à l'aide de données réelles concernant des variables de santé. Une telle évaluation permettrait d'établir si les EPR fonctionnent bien dans des situations réelles, s'ils sont biaisés, précis, faciles à utiliser, etc.

Nous avons donc décidé d'entreprendre cette étude d'évaluation, dans le but d'en connaître davantage quant à l'exactitude de certains EPR en santé.

3. Méthode

Les données de certaines enquêtes nationales seront combinées à des informations auxiliaires locales pour produire des estimés pour petites régions. Ces estimés locaux seront alors comparés à la "vérité" à l'échelle de la petite région, cette "vérité" provenant d'enquêtes locales en santé réalisées dans la petite région. Différentes mesures de "distance" seront utilisées lors des comparaisons: variation et variation moyenne due à l'erreur, racine carrée de la variation moyenne due à l'erreur, corrélations linéaires, etc.

Certains DSC de la province de Québec (représentant, en moyenne, 200 000 individus) constitueront les petites régions étudiées et, à une échelle encore plus petite, les CLSC qui composent ces DSC (représentant, en moyenne, 40 000 individus).

L'étude se déroulera en deux phases. Dans une première phase, les estimés pour petites régions seront produits à partir des données de l'enquête Santé Canada (1978-1979) et comparées aux données collectées au DSC de l'Hôpital Général de Montréal (HGM) lors de l'enquête Santé Communautaire (1979-1981). Dans une seconde phase, des estimés seront produits à l'aide des données de l'enquête nationale Promotion de la Santé (1985) et comparés aux données obtenues lors de la répétition de cette enquête à Montréal, Laval et Hull (1985-1986) (un total de 5 DSC).

---

[1] Sur huit études d'évaluation empirique dans le domaine de la santé, presque toutes concernent l'estimateur synthétique et toutes mentionnent l'importance de répéter ces évaluations.

L'enquête Santé Canada concerne 32 000 Canadiens (dont 6 000 Québécois) interrogés quant à leurs habitudes de vie, leur état de santé et l'utilisation qu'ils font des services médicaux. Seule la portion québécoise de cette enquête sera utilisée pour produire les estimés. L'enquête Santé Communautaire est une répétition de l'enquête Santé Canada au DSC-HGM, sur un échantillon de 1300 individus membres de cette communauté.

L'enquête Promotion de la Santé touche 11 000 Canadiens, dont 850 Québécois, qui furent interrogés concernant leur état de santé, ainsi que leurs connaissances, comportements, intentions, attitudes et croyances quant à la santé. Il s'agit donc d'une enquête s'intéressant à des aspects plus mous de la santé. Deux séries d'estimés pour petites régions seront produites: une à l'aide de la portion québécoise de l'enquête seulement et l'autre à l'aide de l'information concernant le total des 11 000 Canadiens. Ces deux séries nous permettront d'observer l'effet de l'échange entre une précision accrue grâce à un plus gros échantillon (11 000 Canadiens) et une augmentation de la corrélation entre les variables auxiliaires et la variable d'intérêt quand les données proviennent d'une zone plus serrée autour de la petite région (850 Québécois).

Les étapes de l'étude seront les suivantes:

- Choix des variables d'intérêt quant à leur utilité, précision et relation avec les autres variables disponibles.

- Choix des EPR quant à leur simplicité et leur "utilisabilité" compte tenu des informations disponibles.

- Calcul des estimés.

- Comparaison des estimés avec la "vérité".

- Conclusions et recommendations.

4. Exemples d'estimateurs pour petites régions

Supposons que nous désirions en connaître davantage au sujet d'une variable "Y" dans une petite région "r". Par exemple, "Y" pourrait représenter une mesure de l'activité sportive des individus "i" de notre petite région. "s" référerait aux différentes strates socio-démographiques (âge-sexe-sse).

4.1 Estimateur simple direct

Quand les seules données disponibles sont celle nous provenant d'une enquête de santé effectuée dans une vaste région, la façon la plus directe de caractériser l'activité sportive des individus d'une petite région incluse dans la vaste région en question consiste à utiliser l'estimateur simple direct. Il s'agit de faire une moyenne de l'activité sportive des quelques individus de la

petite région qui ont été échantillonnés lors de la vaste enquête. La plupart du temps cependant, le nombre de sujets touchés sera trop petit pour permettre l'obtention d'un estimé précis.

4.2 Estimateur synthétique

Cet estimateur correspond à la standardisation indirecte si familière aux épidémiologistes.

$$\hat{Y}_{r..} = \Sigma_s (Y_{.s.} \times P_{rs.})$$

$Y_{.s.}$ = activité sportive des individus de la strate "s" (dans l'enquête nationale)

$P_{rs.}$ = "importance" de la strate "s" dans la petite région

Il s'agit de faire une somme pondérée des valeurs nationales de l'activité sportive "Y" pour chaque strate "s". Les poids correspondent à l'"importance" d'une strate "s" particulière (par rapport aux autres strates) dans la petite région d'intérêt. La condition sous-jacente à cette estimation est que le niveau d'activité sportive dans une strate "s" particulière à l'échelle nationale est le même que celui de la même strate au niveau local.

4.3 Estimateur de régression

Cet estimateur consiste à calculer une équation de régression de la relation entre la variable d'intérêt (activité sportive) et certaines variables auxiliaires que l'on pense associées à l'activité sportive.

Construction de l'équation de régression:

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\beta}_2 W_i + \hat{\beta}_3 Z_i$$

Y_i = activité sportive pour chaque individu de l'enquête nationale (non-disponible à l'échelle de la petite région)

X_i, W_i, Z_i = variables auxiliaires corrélées à "Y" pour chaque individu échantillonné dans l'enquête nationale (disponible à l'échelle de la petite région)

L'information concernant la variable d'intérêt "Y" n'est disponible qu'à l'échelle nationale, tandis que les variables auxiliaires le sont aussi à l'échelle locale. Une fois l'équation construite et les coefficients de régression déterminés, on l'utilise afin de prédire l'activité sportive dans la petite région d'intérêt, en remplaçant les variables auxiliaires par leur valeur moyenne dans la petite région.

Application de l'équation de régression:

$$\hat{Y}_{r..} = \hat{\beta}_0 + \hat{\beta}_1 \overline{X}_{r..} + \hat{\beta}_2 \overline{W}_{r..} + \hat{\beta}_3 \overline{Z}_{r..}$$

(pour une équation linéaire)

Un des avantages de cette méthode est sa grande flexibilité quant au type de variables auxiliaires qui peuvent être utilisées. En effet, le modèle se prête aussi bien à l'utilisation de variables continues que catégoriques. Il permet même l'utilisation de variables écologiques.

4.4 Estimateur composé

L'estimateur composé consiste à faire une somme pondérée de deux ou plusieurs estimateurs, le but étant d'optimiser les poids de façon à obtenir une variation dûe à l'erreur plus petite que celle d'une des composantes et parfois même plus petite que celle de toutes les composantes.

$$\hat{Y}_{r..} = \alpha_1 \hat{Y}sd_{r..} + \alpha_2 \hat{Y}s_{r..}$$

Il existe toute une littérature concernant les poids à donner à de tels estimateurs et nous devrons faire un choix à ce niveau.

5. Problèmes anticipés

Bien que le dispositif de ce travail soit simple et direct, nous anticipons quelques problèmes méthodologiques que nous nous contenterons de mentionner ici. Tout d'abord, la "vérité" contre laquelle seront comparés nos estimés n'en constitue pas vraiment une. Il s'agit plutôt d'un estimé d'une précision supérieure à ceux produits par l'enquête nationale et qui ne devrait pas être biaisé, au moins en ce qui concerne la totalité des DSC. De plus, comme cette "vérité" provient d'une enquête qui n'a pas été menée tout à fait simultanément à l'enquête nationale, on peut s'attendre à des différences non-explicables entre les estimés directs et les estimés produits à l'aide des EPR, particulièrement en ce qui a trait aux variables les plus instables.

Les enquêtes nationales que nous utiliseront ont aussi des limites. En effet, les données disponibles sur les rubans magnétiques distribués aux utilisateurs ne correspondent pas aux données brutes. Il s'agit de données transformées et pondérées de

façon à tenir compte de l'effet des non-répondants. De plus, par souci de confidentialité, certaines informations ont été éliminées de ces rubans. Une partie de ces données nous auraient permis d'utiliser d'autres EPR sans toutefois nous permettre l'identification des individus.

Finalement, la "distance" qui sera calculée entre les estimés et la "vérité" ne sera pas toujours facile à interpréter. Il s'agira toujours d'une combinaison entre un biais et une erreur aléatoire, mais, dans certains cas, il nous sera impossible ou très difficile de séparer ces deux sources d'erreur.

En résumé, ce travail consistera en une évaluation de certains estimateurs pour petites régions que nous choisirons en fonction des données qui nous sont accessibles. Des estimés de certains paramètres de santé seront produits, à l'aide de ces EPR, pour certains DSC et CLSC du Québec et comparés aux "vraies" valeurs pour ces mêmes régions, elles-mêmes étant des estimés d'une plus grande précision et théoriquement non-biaisés. Le comportement des estimateurs en fonction des différents types de variables prédites permettra d'établir les conditions optimales d'utilisation de chacun et leur performance dans différentes circonstances.

SEGMAG: Analyse de Données par Construction d'Arbres en Biostatistique

Antonio Ciampi, Johanne Thiffault et Roxane du Berger
Département d'Epidémiologie et Biostatistiques, Université McGill
Montréal, Québec

Nous montrons ici que SEGMAG, une méthode pour la construction d'arbres à partir de données, est spécialement adaptée au traitement des problèmes épidémiologiques, grâce à la généralité du critère et à la possibilité d'effectuer des analyses par sous-groupes. Nous présentons aussi plusieurs exemples d'intérêt clinique.

MOTS-CLES: Données censurées; Variables déterminantes; Variables de confusion; Variables modificatrices; Modèle de Régression Généralisé.

1. INTRODUCTION

L'informatisation des données médicales est une réalité de plus en plus présente, voire envahissante, à un point tel que des congrès entiers sont consacrés à l'étude des problèmes de confidentialité qui en sont la conséquence inévitable.

Ce qui est peut-être moins évident, aux yeux du grand public, est que si d'une part l'informatisation des données médicales pose des problèmes sérieux, elle offre des avantages qui dépassent largement le cadre administratif. En effet, elle peut constituer une opportunité d'accroître la compréhension scientifique de certaines maladies et de la santé en général. En fait, les bases de données médicales pourraient être comparées, à juste titre, à une mine de connaissances qui est encore loin d'être dûment exploitée.

Une des raisons de la sous-utilisation des données médicales à des fins scientifiques est l'absence de méthodes d'analyse appropriées. L'étude des données médicales, colligées systématiquement mais sans but scientifique précis, nécessite des méthodes dont la clarté descriptive serait égale à leur pouvoir de généralisation.

Dans cet article, nous présentons une adaptation à la Biostatistique d'une méthode d'analyse exploratoire bien connue dans d'autres domaines d'application, soit la construction d'arbres de classification à partir d'un ensemble de données (Breiman et al. 1984; Sonquist et al. 1964).

La structure d'arbre constitue un moyen naturel et économique de résumer le contenu informationnel d'un ensemble de données. Les noeuds y sont définis par de simples questions concernant certaines caractéristiques d'un individu, tandis que les feuilles (noeuds terminaux) correspondent à des profils distincts. On retrouve la structure d'arbre dans l'organisation de tout catalogue, entre autres dans les systèmes de classification utilisés dans les bibliothèques. Elle représente également l'outil privilégié de l'informaticien pour organiser la mémoire d'un ordinateur.

Fait encore plus pertinent, le branchement récursif décrit par une structure d'arbre ressemble beaucoup au raisonnement scientifique et clinique, qui procède par élimination successive ou par l'enchaînement de simples pas logiques visant à l'identification d'un patron global et généralisable.

Malgré ces caractéristiques fort avantageuses, l'utilisation des arbres de classification n'est pas très répandue dans les sciences de la santé. Bien que quelques adaptations de la méthode proposée par Sonquist et al. (1964) aient déjà été publiées dans la littérature (Segal, 1988; Ciampi et al. 1987, 1988a, 1988b), aucune tentative systématique n'a été entreprise afin d'établir le lien entre, d'une part les problèmes que se posent les biostatisticiens en analysant des données médicales, et d'autre part les méthodes de construction d'arbres. C'est ce que nous essaierons de faire ici en ce qui a trait au domaine particulier de l'Epidémiologie.

La section 2 de ce travail est consacrée à la description du contexte de l'analyse des données en Epidémiologie, ainsi qu'aux modèles de régression, qui en constituent les méthodes analytiques courantes. La méthode de construction d'arbres SEGMAG (SEGMentation et AGrégation) est présentée dans la section 3. Des applications épidémiologiques sont exposées dans la section 4 qui est suivie par la conclusion.

2. DONNEES, MODELES DE REGRESSION ET PROBLEMES D'ANALYSE EN EPIDEMIOLOGIE

Les données épidémiologiques se présentent habituellement sous la forme classique d'une matrice dont les lignes représentent les individus et les colonnes les variables. Il faut noter toutefois qu'un cas peut être soit une personne, soit une unité observationnelle plus complexe, telle une usine, une ville ou une région, dans laquelle on étudierait des événements concernant la santé ou la maladie.

Les variables mesurées au cours d'une étude épidémiologique sont classifiées comme variables-réponse, variables déterminantes, confondantes et modificatrices.

Les variables-réponse sont celles ayant un lien direct avec l'événement qui constitue l'objet de l'étude. Elles sont de types variés et différents des variables quantitatives et continues qu'on retrouve dans les problèmes classiques. Les variables-réponse typiques en épidémiologie sont les variables de dénombrement (nombre de cas d'une certaine maladie dans une région donnée, nombre de rechutes chez un malade, etc.), les variables qualitatives, qu'elles soient nominales (maladies, complications, etc.) ou ordonnées (degré d'amélioration suivant une intervention, degré de risque de développer une maladie, etc.), ou encore des variables connues exactement pour certains individus, mais connues seulement en partie pour d'autres (temps censurés ou tronqués).

Les variables déterminantes, ou déterminants, sont celles dont l'effet sur la réponse est d'un intérêt particulier dans l'étude en question. Il peut s'agir de variables observées sans intervention directe, telles que l'âge d'un patient ou le revenu moyen d'une certaine région. D'ailleurs ce sont souvent des variables dont les valeurs sont contrôlées, telles que les variables décrivant des régimes de traitement.

Les variables confondantes sont des variables qui sont observées pour tenir compte de leurs effets, connus ou soupçonnés, susceptibles de masquer l'effet des déterminants sur la réponse. Une variable confondante est donc associée en même temps à la réponse et aux variables déterminantes, ce qui cause des problèmes d'analyse et d'interprétation bien connus.

Les variables modificatrices, ou modificateurs, sont des variables susceptibles d'altérer l'effet d'un déterminant sur la variable-réponse.

Par exemple, un certain médicament peut être très efficace pour le traitement d'une infection, sauf pour les individus qui manquent d'un certain enzyme et pour lesquels il devient extrêmement toxique. Dans ce cas, la variable indicatrice du manque d'enzyme modifie l'effet du traitement, et est donc considérée comme une variable modificatrice.

Comme dans tout autre domaine d'application, les modèles de régression occupent une place centrale dans l'analyse de données épidémiologiques. Chaque type de variables-réponse particulier a stimulé, par ailleurs, l'élargissement du modèle linéaire classique pour englober des situations plus générales. La régression ordinaire a été généralisée au cas d'une réponse continue multivariée, alors que la régression logistique a été introduite pour traiter des réponses qualitatives (Cox 1970). De même, la régression de Poisson et ses extensions (Lawless 1987) sont employées quand la variable-réponse est une variable de dénombrement, alors que les régressions exponentielle, de Weibull ou de Cox sont utilisées couramment pour traiter les données censurées. La plupart de ces méthodes représentent des cas particuliers de ce qu'on appelle le modèle linéaire généralisé (McCullagh et Nelder 1983), popularisé par le logiciel GLIM. De plus en plus, ce modèle constitue l'outil de base des analystes de données épidémiologiques.

Les problèmes analytiques les plus fréquents en épidémiologie sont les problèmes de prédiction et les analyses par sous-groupes. Les premiers consistent à prédire la réponse à partir des déterminants, tout en corrigeant pour l'effet des variables de confusion. Les analyses par sous-groupes, pour leur part, visent à déceler des effets de modification.

Deux solutions générales ont été proposées pour corriger l'effet des variables confondantes aux fins de démasquer celui d'un déterminant. La première, proposée par Miettinen (1976), consiste à former un score de confusion, à savoir une combinaison linéaire des confondants obtenue par des méthodes de régression classiques, pour ensuite étudier l'effet des déterminants sur des sous-populations d'individus ayant des scores de confusion comparables.

L'autre solution a été proposée par Rosenbaum et Rubin (1984), dans le cas particulier où on a un unique déterminant qui est une variable indicatrice de traitement. Ces auteurs proposent de déterminer aussi une combinaison linéaire des confondants, mais dans le but de prédire, non pas la réponse, mais plutôt la probabilité d'affectation à un régime de traitement donné; ils appellent cette combinaison un score de propensité, et ils montrent qu'en corrigeant pour la propensité, on élimine une plus grande proportion du biais de confusion que par toute autre méthode.

Les deux solutions proposées se ressemblent en ce que l'analyste cherche d'abord une stratification, pour tester ensuite l'effet des déterminants selon les strata.

Dans le cas de l'analyse par sous-groupes, qui vise à déceler des effets de modification, on cherche à déterminer, soit des groupes de patients pour lesquels un remède est contre-indiqué, soit des groupes d'individus pour lesquels un traitement est spécialement efficace. Cette recherche se fait souvent en testant la présence d'interactions entre les déterminants et les variables qui pourraient être modificatrices.

3. SEGMAG: FORMULATION GENERALE ET CHOIX DE BASE

On a déjà observé que l'analyse des données épidémiologiques utilise la stratification comme pas intermédiaire, puisque l'étude de l'effet d'un déterminant requiert souvent qu'on partitionne préalablement les données en strata homogènes par rapport soit au score de confusion, soit au score de propensité.

De façon générale, on pourrait dire que le but d'une analyse épidémiologique est souvent de déterminer des sous-populations qui sont homogènes dans un sens précis. Les problèmes de prédiction, par exemple, font plus souvent appel à l'identification de sous-groupes à pronostique homogène, qu'à une estimation précise basée sur les variables. Par ailleurs, l'analyse par sous-groupes vise directement l'identification de sous-populations pour lesquelles l'effet d'un déterminant soit homogène.

Or, les méthodes traditionnelles de régression permettent d'obtenir des strata de façon simple et très générale. Il suffit de produire un modèle linéaire à partir des variables déterminantes et de calculer le score pour chaque individu. On détermine ensuite des intervalles de score correspondant à la stratification désirée. Cette approche est pourtant limitée en ce que les sous-groupes ainsi obtenus sont difficiles à décrire d'une manière directe. De plus, la définition des intervalles à partir du score ne peut se faire que par des choix arbitraires (choix du nombre d'intervalles, ou des bornes).

Le but de SEGMAG est d'obtenir des classes, homogènes et distinctes dans un sens spécifié, et décrites avec la simplicité qui caractérise la structure d'arbre. Bien que SEGMAG soit basé sur des modèles de régression généralisés, l'approche à la modélisation y est radicalement différente; au lieu de chercher une combinaison linéaire de variables, SEGMAG cherche une structure d'arbre. Cette structure définit des variables indicatrices de classe, qui elles seront les constituantes d'un modèle de régression.

Dans l'approche SEGMAG on distingue deux types de variables: variables-critère et de prédiction. Le but de l'analyse est de construire un arbre dont les noeuds terminaux représentent des classes homogènes et distinctes du point de vue d'un paramètre, nommé critère, qui est estimé à partir des variables-critère. Les branchements successifs de l'arbre se font en répondant à des questions concernant les variables de prédiction.

Ce type de classification est relié à la classification épidémiologique introduite ci-haut de la façon suivante. Les variables-critère peuvent être, soit des variables-réponse seulement, soit une combinaison de variables-réponse, variables déterminantes et variables confondantes. Dans le premier cas, on parle de classification prédictive sans variables de confusion, où le critère est un paramètre caractérisant la distribution de la variable-réponse. Dans le deuxième cas il s'agit, soit de classification prédictive avec variables de confusion (le critère étant défini de la même façon mais avec, en plus, des paramètres nuisibles permettant de corriger pour l'effet des variables de confusion), soit d'analyse par sous-groupes (le critère étant alors défini comme le coefficient de la régression des variables-réponse sur les déterminants).

Quant aux variables de prédiction, elles peuvent être soit des déterminants (classification prédictive), soit des variables confondantes (classification pour obtenir des groupes homogènes du point de vue des effets de confusion ou du point de vue de la propensité), soit des variables modificatrices (analyse par sous-groupes).

Outre les choix préliminaires dépendant du but de l'analyse (spécification des variables-critère, du critère comme tel, et des variables de prédiction), l'utilisateur de SEGMAG devra faire trois choix fondamentaux, nécessaires pour définir l'algorithme. Il devra donc choisir (a) la mesure de dissimilarité pour comparer deux populations, (b) la famille de questions définissant les branchements, (c) la règle d'arrêt pour les algorithmes d'émondage et d'agrégation qui seront décrits à la fin de la présente section.

Un choix approprié de la famille de questions permettra d'obtenir des résultats d'interprétation facile. Le choix le plus commun est celui des questions simples, de la forme "Est-ce que [z>a] ?", pour tout a faisant partie du domaine de z, dans le cas où z est un prédicteur ordonné; pour les prédicteurs nominaux on considère les questions de la forme "Est-ce que [z∈A]?", pour tout sous-ensemble A du domaine de z. D'autres choix sont possibles, dont les combinaisons booléennes de questions simples et les listes de propriétés.

Ces choix faits, la construction des classes par SEGMAG procède en trois étapes: (a) construction d'un grand arbre (SEGMentation), (b) émondage du grand arbre pour obtenir un arbre "honnête", (c) AGrégation des feuilles de l'arbre honnête pour obtenir des classes distinctes.

La première étape débute par une partition binaire, opérée sur la population entière, de manière à obtenir deux sous-populations. Parmi toutes les façons possibles de partitionner la population de départ, on choisit celle qui a le plus grand contenu informationnel par rapport au critère (à l'intérieur de la famille de questions choisie). On continue ainsi la segmentation de manière itérative, en opérant sur les sous-populations obtenues à l'étape précédente, et ce jusqu'à l'obtention de noeuds terminaux assez petits, dont la taille est choisie à l'avance par l'utilisateur (dans les exemples de cet article la taille minimale est fixée à 10). Le contenu informationnel est défini selon la mesure de dissimilarité, qui est dans la plupart des cas la statistique du rapport de vraisemblance (SRV) de l'hypothèse que le critère est différent d'une sous-population à l'autre, contre l'hypothèse qu'il ne diffère pas.

La SRV est basée sur le modèle de régression généralisé suivant:

$$\theta = \beta_0 + \beta_1 \mathbf{x} + \beta_2 \mathbf{u} + \gamma \mathbf{I} + \eta \mathbf{I}\mathbf{x} \qquad (1)$$

où les lettres grecques représentent des paramètres à estimer, $\mathbf{x}$ dénote le vecteur des déterminants, $\mathbf{u}$ les variables de confusion et $\mathbf{I}$ la variable indicatrice de la partition. Dans le cas de la classification prédictive, β_1 est égal à zéro et l'hypothèse considérée est que $\gamma = 0$ (γ est le critère). Dans le cas d'une analyse par sous-groupes, on teste l'hypothèse $\eta = 0$ (η est le critère).

L'émondage du grand arbre a pour but le choix d'un arbre moins détaillé mais plus généralisable à d'autres échantillons, d'où l'appellation d'arbre "honnête", qu'on retrouve souvent dans la littérature (Breiman et al. 1984). Les éléments essentiels de ce choix sont: (a) la construction d'une chaîne de sous-arbres à partir du grand arbre, (b) l'estimation, pour chaque sous-arbre, d'un modèle de régression généralisant d'une façon naturelle celui de l'équation (1), et (c) le choix final du sous-arbre satisfaisant un certain critère d'optimalité. Dans les exemples qui suivront, ce critère de choix est basé sur le critère informationnel d'Akaike (AIC).

L'agrégation est une étape qu'on ne retrouve pas dans les autres algorithmes de construction d'arbres et qui consiste à joindre ensemble

les feuilles (noeuds terminaux) de l'arbre honnête pour construire une chaîne de modèles, parmi lesquels on choisira le modèle final, correspondant aux classes SEGMAG.

4. EXEMPLES D'APPLICATIONS

Afin d'illustrer les idées énoncées précédemment, nous présentons quatre exemples d'analyse de données épidémiologiques, chacun mettant en évidence une des applications possibles de la méthode SEGMAG. Dans un premier temps, nous présenterons les résultats d'une analyse concernant le développement neuro-psychologique chez l'enfant ayant été victime d'une méningite. Cet exemple nous permettra de démontrer la possibilité d'utiliser SEGMAG comme méthode de prédiction d'un construit multivarié, puisque le développement neuro-psychologique est défini ici en termes de deux indices (facteurs) obtenus à l'aide d'une Analyse Factorielle. Le deuxième exemple illustre également l'originalité de SEGMAG, puisqu'on y présente une analyse dont l'objectif principal est la prédiction d'une probabilité. Bien que les modèles de régression logistique représentent une alternative bien connue, la méthode SEGMAG offre le double avantage d'une interprétation clinique plus aisée et d'une prise en considération automatique des interactions entre les variables de prédiction. Le troisième exemple est une démonstration de SEGMAG comme méthode permettant d'élaborer la typologie d'une maladie. Les résultats d'une analyse exploratoire sur le diabète insulino-dépendant seront présentés à titre d'illustration de la stratégie d'analyse que nous proposons pour résoudre ce problème. Enfin, le dernier exemple est tiré d'une analyse concernant le cancer du poumon, et servira ici à démontrer la possibilité de traiter les différents types de variables décrits à la section précédente, soient les variables de confusion, de modification et de prédiction.

4.1 Développement neuro-psychologique à la suite d'une méningite

Avant de décrire la démarche méthodologique utilisée, il convient d'exposer brièvement quelques détails concernant les données qui ont été recueillies. On sait déjà que l'enfant ayant subi une méningite en bas âge a un risque relativement élevé de présenter des séquelles par la suite, sous la forme d'un retard plus ou moins marqué dans son apprentissage verbal ou non-verbal. La prédiction de ces séquelles, à l'aide de caractéristiques présentes au moment du diagnostic de la méningite, représente un atout important dans le traitement de la maladie et de ses conséquences sur le développement de l'enfant. En effet, toute démarche permettant de détecter de façon plus rapide les anomalies d'apprentissage résultant de la méningite, permettrait aux intervenants d'entreprendre plus tôt les mesures correctives nécessaires, et ainsi de diminuer l'impact de la maladie sur la santé mentale de l'enfant.

Les données, recueillies entre 1985 et 1988 dans trois hôpitaux du Québec et de l'Ontario, constituent tous les enfants ayant été victimes d'une méningite et dont la famille compte au moins un autre enfant pouvant servir de témoin dans l'étude. Les variables observées sont de deux types: variables de prédiction, observées au moment du diagnostic, et variables-critère, à partir desquelles on mesurera le développement neuro-psychologique. Parmi les variables de prédiction, on compte par exemple la durée des symptômes, le décompte des globules blancs, la

durée de la fièvre, etc. Pour mesurer le développement neuro-psychologique, on a administré aux enfants une batterie de tests, dont les résultats ont été soumis à une analyse factorielle, ce qui a permis de dégager deux indices de développement neuro- psychologique, le premier qu'on a interprété comme un indice d'apprentissage verbal, F1, et le deuxième comme un indice d'apprentissage non-verbal, F2.

Après avoir défini les variables de prédiction et les variables-critère, on doit déterminer le critère comme tel, qui sera défini ici comme étant l'espérance de F1 et F2. Finalement, la dissimilarité sera définie comme le rapport de vraisemblance permettant de tester l'hypothèse d'égalité des moyennes, contre l'hypothèse d'inégalité; on suppose ici que F1 et F2 suivent une loi normale bivariée.

Bien que l'échantillon soit de petite taille, la méthode permet une analyse exploratoire dont les résultats s'interprètent aisément de façon clinique, comme le démontre la Figure 1.

..... Figure 1 .....

A titre d'exemple d'interprétation, considérons le groupe de patients dont les séquelles sont plus marquées, soit le groupe de 25 patients qu'on retrouve au bas de l'arbre à droite. Précisons d'abord que ce noeud terminal résulte de l'agrégation de deux noeuds obtenus lors de la partie segmentation, ce qui indique que ces deux groupes de 14 et 11 patients, respectivement, ont des indices d'apprentissage (verbaux et non-verbaux) semblables. Si on tente de caractériser ces patients en suivant le cheminement de l'arbre, de la racine à ce noeud terminal, on constate que l'évolution de la maladie se manifeste de la façon suivante. Ces enfants présentaient, lors du diagnostic de méningite, des taux de globules blancs inférieurs à la normale, ce qui indique que leur système immunitaire n'avait pas répondu adéquatement à l'attaque de la maladie. Au deuxième palier de l'arbre, on voit que ces enfants avaient été victimes d'une attaque rapide, donc plus agressive de la maladie, comme l'indique la durée des symptômes qui est de moins de 24 heures. Enfin, pour ce qui est du groupe de 14 patients qu'on retrouve au troisième palier de l'arbre, et qui font aussi partie du groupe de 25 qui nous intéresse, l'interprétation est d'autant plus aisée qu'un niveau élevé de protéines dans le liquide céphalo-rachidien est un indice d'agressivité de la maladie. Quant au groupe de 11 patients qu'on retrouve au dernier palier de l'arbre, plusieurs interprétations sont possibles pour expliquer leur présence parmi le groupe dont les séquelles sont les plus sévères. L'une d'elles serait qu'une fièvre de courte durée est un indice d'une réponse immunitaire inadéquate; cependant la petite taille de l'échantillon ne nous permet pas d'affirmer de façon définitive qu'une telle interprétation soit plausible, puisque cette partition peut très bien être due à des variations aléatoires de l'échantillon.

4.2 Prédiction de la probabilité de chute chez les personnes âgées

Dans cette analyse, la population visée constitue les 220,000 personnes de 65 ans et plus vivant dans le district avoisinant de l'Hôpital Général de Montréal. Parmi ce groupe, on en compte environ 2,800 qui vivent continuellement à la maison mais qui ont régulièrement besoin de soins infirmiers. La majorité de ces personnes correspondent à une catégorie de gens qu'on pourrait qualifier de fragiles, c'est-à-dire qu'ils souffrent, soit de pathologies multiples, soit d'une diminution de leur capacité de fonctionnement, soit d'une diminution de mobilité, ou encore d'une perte de lucidité. Ces patients étant plus susceptibles de faire des chutes, on considère important d'en prédire la probabilité, afin de permettre aux intervenants de décider du suivi approprié (visites à domicile, hospitalisation, co-habitation, etc.).

On a recueilli des données auprès de 277 personnes âgées et on a retenu 7 facteurs de risque associés à la probabilité de faire une chute, soit : (a) les étourdissements, (b) une démarche anormale, (c) l'urgence ou la fréquence urinaire, (d) un accident cérébro-vasculaire au cours des 5 dernières années, (e) l'utilisation d'une canne ou d'une marchette, (f) une déficience quelconque au niveau des membres inférieurs, et (g) l'arythmie cardiaque.

Puisque le but de l'analyse était de prédire la probabilité de chute, on a d'abord tenté de développer un modèle de régression logistique, mais on a observé la présence d'interactions parmi les variables de prédiction. Plutôt que d'introduire artificiellement un certain nombre d'interactions, on a choisi d'utiliser SEGMAG pour obtenir ce qu'on pourrait appeler un arbre de régression logistique,

dont l'interprétation clinique est beaucoup plus aisée et qui tient compte automatiquement des interactions entre les variables de prédiction.

Les variables de prédiction sont définies ici comme étant les 7 facteurs de risque énoncés plus haut, et la variable-critère est une variable binaire qui indique si, oui ou non, le patient a fait une chute lors des 12 derniers mois. Finalement, le critère comme tel est défini comme la probabilité de faire une chute.

Comme on peut le voir à la Figure 2, l'arbre de régression fournit au professionnel de la santé un outil permettant de déterminer facilement et rapidement le risque de chute pour chaque patient. En effet, les variables de prédiction qu'on retrouve dans le modèle final peuvent être facilement observées (dans le cas de la démarche) ou obtenues en questionnant le patient afin de déterminer s'il souffre d'étourdissements ou de fréquence/urgence urinaire. Sous chaque noeud terminal, on retrouve la probabilité de faire une chute, et en suivant le cheminement qui mène de la racine de l'arbre à ce noeud terminal, on peut facilement déterminer le profil correspondant.

..... insérer ici la Figure 2 .....

Par exemple, le groupe de 76 patients qu'on retrouve à droite au dernier palier de l'arbre correspond à un groupe à risque élevé, leur probabilité de faire une chute étant de 0.75. Le profil de ces personnes se caractérise par des étourdissements et des problèmes de fréquence/ urgence urinaire. Le principal avantage de la méthode SEGMAG, dans ce cas-ci, est que l'outil de prédiction qu'on obtient est simple à utiliser et ne fait pas appel à la mémoire du patient, puisqu'il ne dépend que de l'observation directe et de quelques questions auxquelles il peut répondre facilement.

4.3 Diabète Insulino-Dépendant

Le troisième exemple est tiré d'une étude longitudinale présentement en cours à l'Hôpital de Montréal pour Enfants, ayant comme principal objectif la détermination du processus qui engendre la destruction progressive des cellules β chez les enfants atteints du diabète insulino-dépendant (IDDM). Cependant l'objectif que nous poursuivons dans notre analyse statistique est d'un autre niveau, puisque nous tenterons de déterminer la typologie du diabète insulino-dépendant. Le problème qui nous intéresse plus particulièrement est donc de savoir si le diabète insulino-dépendant existe sous différents sous-types, ou si au contraire il est de type unique. La nature de cette analyse, qui se veut purement exploratoire, ne nous permet pas de formuler des hypothèses quant à l'étiologie de la maladie. Cependant elle représente une première étape essentielle qui pourrait mener à l'élaboration d'une typologie du diabète insulino-dépendant.

..... Figure 3 .....

Les variables observées sont à la fois nombreuses et de types variés (continues, discrètes et binaires), ce qui représente une difficulté de l'analyse. Pour remédier à ce problème, on a adopté la stratégie consistant à réduire le nombre de variables par l'Analyse des Correspondances, pour ensuite procéder à une classification des patients à partir des facteurs obtenus à l'étape précédente.

La première étape de l'analyse, soit la réduction des variables, a permis de dégager 5 facteurs expliquant 75% de l'inertie. L'étape suivante, soit la classification par la méthode SEGMAG, utilise ces facteurs comme variables-critère, le but étant de former des groupes de patients dont la distribution des facteurs est homogène. Le critère est défini comme l'espérance de ces facteurs, dont on suppose qu'ils suivent une loi normale multivariée, et les variables de prédiction constituent toutes les variables originales.

L'arbre illustré à la Figure 3 semble suggérer la présence de trois sous-types de diabète. D'une part, on observe un groupe de 26 patients qui souffrent d'un diabète plutôt agressif, comme l'indique la présence de coma. A l'autre extrémité de l'arbre, on a un groupe de 58 patients qui semblent être atteints d'une forme de diabète moins sévère, comme l'indiquent l'absence de coma et la production normale de C-peptides. Le groupe intermédiaire, composé de 53 patients, est caractérisé par l'absence de coma, mais par une diminution marquée de la sécrétion de C-peptides par les cellules β.

Table 1. Description des Variables

Variable	Type
Démographiques	
Groupe d'âge	Discrète
Sexe	Binaire
Antécédents familiaux	
Diabète de type IDDM	Binaire
Autres types de diabète	Binaire
Marqueur génétique DR	Discrète
Fonctionnement des cellules β	
Production de C-peptides	Continue
Besoins en Insuline	Continue
Arrêt de fonction	Binaire
Activité auto-immune dirigée contre ...	
Les ilôts pancréatiques	Binaire
L'insuline	Binaire
La glande thyroïde	Binaire
Système Endocrinien	
T4	Continue
TSH	Continue
Cortisol	Continue
Contrôle du Métabolisme	
HbA1	Continue
Cholestérol	Continue
Triglycérides	Continue
Glucose sanguin	Continue
Présentation Clinique	
Coma	Binaire
pH	Continue
Durée des symptômes	Continue

4.4 Cancer du poumon de type "petites cellules"

Cet exemple est tiré d'une analyse de 614 patients hospitalisés au Princess Margaret Hospital et au Toronto General Hospital, entre 1976 et 1986. Le but principal de l'analyse était d'obtenir une classification pronostique des patients, mais on s'est intéressé plus particulièrement au pouvoir de prédiction du marqueur tumoral LDH.

Les résultats de la classification pronostique ne seront pas illustrés ici mais on peut les trouver dans Ciampi et al. (1988a); nous nous attarderons plutôt à démontrer la possibilité d'effectuer avec SEGMAG une classification de propensité dans le but de corriger l'effet de variables de confusion.

Le marqueur tumoral LDH a été utilisé dans cette étude comme mesure de la progression de la maladie. Etant donné la présence de variables de confusion dans l'analyse, on a procédé d'abord à une classification de confusion, tel que proposé par Miettinen (1976). Pour ce faire, on a supposé que le marqueur LDH suivait une loi multinomiale, et on a tenté de le prédire à l'aide des autres variables. Tel qu'illustré à la Figure 4, on a ainsi obtenu deux groupes d'individus dont la distribution du marqueur LDH est homogène; ces deux groupes, composés de 200 et 88 individus respectivement, serviront de strata à l'étape suivante qui consiste en une régression de Cox visant à prédire la survie à partir du marqueur LDH. Il est à noter que le nombre réduit d'individus utilisés dans cette analyse (288 au lieu de 614) provient du fait qu'on devait éliminer tous les individus pour lesquels la variable LDH était manquante, puisque LDH était utilisée comme variable-réponse à trois niveaux (bas, moyen élevé).

..... Figure 4 .....

Sans nous attarder sur les détails de l'analyse, effectuée avec le logiciel BMDP, nous mentionnerons à titre illustratif les résultats de l'étape suivante, qui consiste en une régression de Cox pour déterminer l'effet de LDH sur la survie. La régression non-stratifiée indique un effet néfaste des valeurs élevées de LDH sur la survie, puisque le risque relatif obtenu est de 1.4762. La régression de Cox stratifiée, qui permet de corriger pour l'effet des variables de confusion, donne un résultat un peu plus conservateur de 1.3290 pour le risque relatif, ce qui confirme tout de même le résultat de l'analyse non-stratifiée.

L'étape suivante consiste à déterminer si l'effet du marqueur LDH est uniforme pour tous les individus, ou si au contraire certains profils des autres variables en modifient l'effet sur la survie. Il s'agit donc ici d'une analyse tenant compte de variables modificatrices, telles que décrites à la section 2. Les résultats de l'analyse sont illustrés dans la Figure 5, qu'on peut interpréter de la façon suivante.

..... Figure 5 .....

Précisons d'abord que la dissimilarité est mesurée en testant, à chaque point de branchement, l'hypothèse selon laquelle le coefficient de régression de LDH sur la survie est le même dans les deux groupes. Une partition s'effectue seulement si le coefficient de régression diffère d'un groupe à l'autre, et les questions définissant la partition (illustrées au centre de chaque branchement) déterminent les variables modificatrices de l'effet de LDH sur la survie.

La principale conclusion qu'on peut tirer en observant cet arbre, est qu'il existe un groupe de 104 individus dont le pouvoir de prédiction de LDH est particulièrement élevé (risque relatif égal à 2.126). Le profil de ces 104 individus se caractérise ainsi: AGE < 65, CEA = bas ou élevé, ALKP = élevé, et BRAI = oui, les symboles étant définis dans la Table 2.

Table 2. Description des Variables

Variable	Description
STAGE	Stade de la maladie
CEA	Marqueur tumoral
BRAIN	Métastases au cerveau
LDH	Marqueur tumoral
MARROW	Examen de la moëlle osseuse
SITES	Nombre de sites de métastases
AGE	Age
ALKP	Phosphatase alcalin
HEMO	Hémoglobine

5. DISCUSSION ET CONCLUSION

La représentation du contenu informationnel d'une matrice de données par un arbre offre un puissant outil de synthèse qui aide à la compréhension et facilite la formulation d'hypothèses, voire de théories scientifiques. Cet outil avait déjà été exploité au cours des 25 dernières années, en particulier par les sociologues (Sonquist et Morgan 1964). Plus récemment, on a considérablement raffiné cet outil en lui ajoutant de la rigueur statistique, par le biais des notions d'arbre honnête et des méthodes de calcul intensif permettant de le construire (Breiman et. 1984). Cependant, ces améliorations se sont jusqu'à présent limitées au contexte du modèle de régression dit ordinaire (une variable-réponse continue et complètement observée pour tous les individus). Cette limitation est probablement un des facteurs qui ont entravé l'utilisation des arbres dans le domaine de la Biostatistique et de l'Epidémiologie.

Dans une série de travaux toujours en phase de développement, nous avons montré que ce puissant outil d'analyse qu'est l'arbre de prédiction peut être adapté aux variables-réponse typiques de l'épidémiologie. Ce que nous proposons est un développement comparable à celui qui a donné lieu au modèle linéaire généralisé, et qui a permis l'analyse de données dans beaucoup de situations typiques de l'épidémiologie. SEGMAG est donc à CART ce que GLIM est aux logiciels standards de régression linéaire ordinaire.

Dans cet article, nous avons voulu souligner le lien entre les capacités de SEGMAG et la problématique qu'on rencontre dans l'analyse de données épidémiologiques.

Les méthodes de construction d'arbre à partir d'un ensemble de données constitue un domaine de recherche où confluent plusieurs disciplines: la Statistique, pour la rigueur inférentielle, les Sciences Informatiques, pour la gestion des calculs, l'Intelligence Artificielle, pour le développement de programmes d'analyse "intelligents", capables d'ajouter aux données, soit des connaissances préalables, soit l'opinion d'experts, et enfin, dans le cas des analyses de données épidémiologiques, les Sciences Biomédicales, pour donner une direction précise aux analyses très flexibles que ces méthodes rendent possibles.

La recherche sur les méthodes de construction d'arbres en est encore à ses premiers pas, et nous sommes convaincus que les prochaines années verront une expansion croissante de cette méthodologie et de méthodologies affines. Cette recherche nous fournira les outils appropriés à la tâche primordiale qui consiste à exploiter à fond la mine d'informations cachées dans les bases de données cliniques.

REFERENCES

Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J. (1984), Classification and regression trees, The Wadsworth Statistics/Probability Series, Belmont, CA.

Ciampi, A., Chang, C.H., Hogg, S.A. and McKinney, S.M. (1987), "Recursive Partition: A versatile method for exploratory data analysis in biostatistics," in Biostatistics (I.B. McNeil and G.J. Umphrey, eds.) D. Reidel Publishing Company, Holland, 23-50.

Ciampi, A., Thiffault, J. et Sagman, U. (1988a), "Evaluation de Classifications par le Critère d'Akaike et la Validation Croisée," Revue de Statistique Appliquée, 36, 33-50.

Ciampi, A., Hogg, S.A., McKinney, S.M. and Thiffault, J. (1988b), "RECPAM: A computer program for recursive partition and amalgamation for censored survival data and other situations frequently occurring in biostatistics. Part I. Methods and program features," Computer Methods and Programs in Biomedicine, 26, 239-256.

Cox, D.R. (1970), The analysis of binary data, London, Chapman Hall.

Lawless, J.F. (1987), "Regression Methods for Poisson Process Data," Journal of the American Statistical Association, 82, 808-815.

McCullagh, P. and Nelder, J.A. (1983), Generalized Linear Models, London, Chapman Hall.

Miettinen, O.S. (1976), "Stratification by a Multivariate Confounder Score," American Journal of Epidemiology, 104, 609-620.

Rosenbaum, P.R. and Rubin, D.B. (1984), "Reducing bias in observational studies using subclassification on the propensity score," Journal of the American Statistical Association, 79, 516-524.

Segal, M.R. (1988), "Regression trees for censored data," Biometrics, 44, 35-47.

Sonquist, J.A. et Morgan, J.V. (1964), The Detection of Interaction Effects, Ann Arbor: Institute for Social Research, University of Michigan.

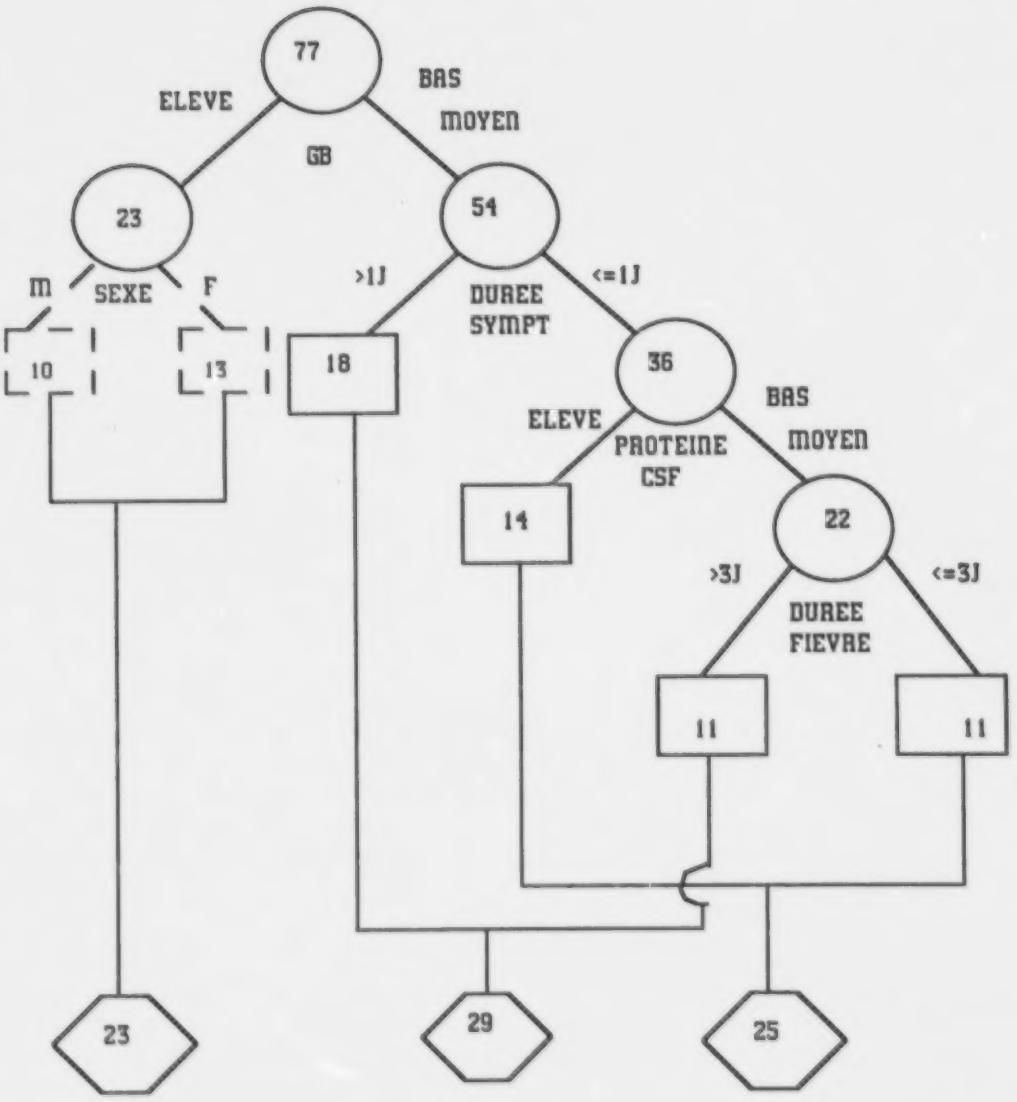
Figure 1. Arbre de prédiction du développement neuro-psychologique

Figure 2. Arbre de prédiction de la probabilité de chute

Figure 3. Arbre de classification du diabète insulino-dépendant

Figure 4. Arbre de prédiction du marqueur LDH (classification de propensité)

Figure 5. Arbre de classification par sous-groupes pour prédire l'effet de LDH sur la survie



$\overline{FACT1}$(SE)= 110(.191) .653(.223) -.855(.229)

$\overline{FACT2}$(SE)= 083(.170) .110(.149) -.520(.220)

Figure 1

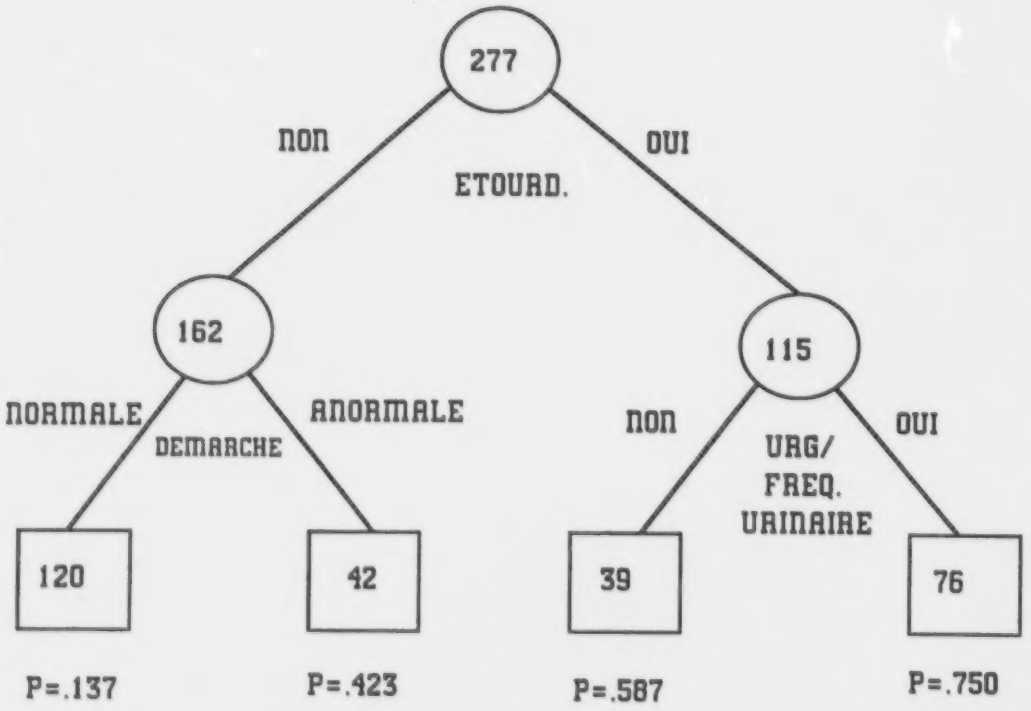

277
NON
OUI
ETOURD.
162
115
NORMALE
ANORMALE
DEMARCHE
NON
OUI
URG/
FREQ.
URINAIRE
120
42
39
76
P=.137
P=.423
P=.587
P=.750

Figure 2

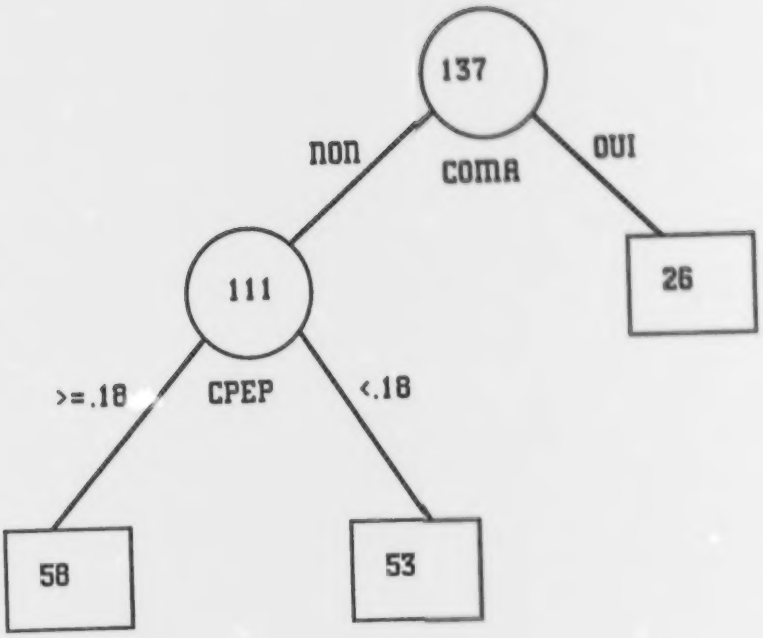


Figure 3

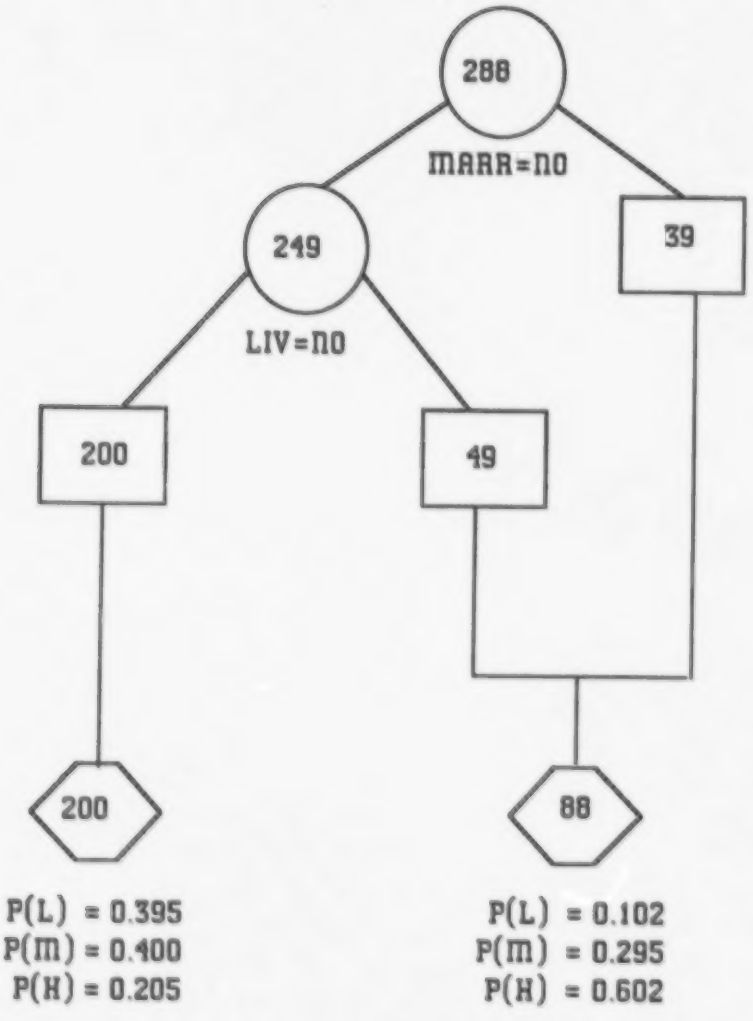

288
MARR=NO
249
39
LIV=NO
200
49
200
88
P(L) = 0.395
P(M) = 0.400
P(H) = 0.205
P(L) = 0.102
P(M) = 0.295
P(H) = 0.602

Figure 4

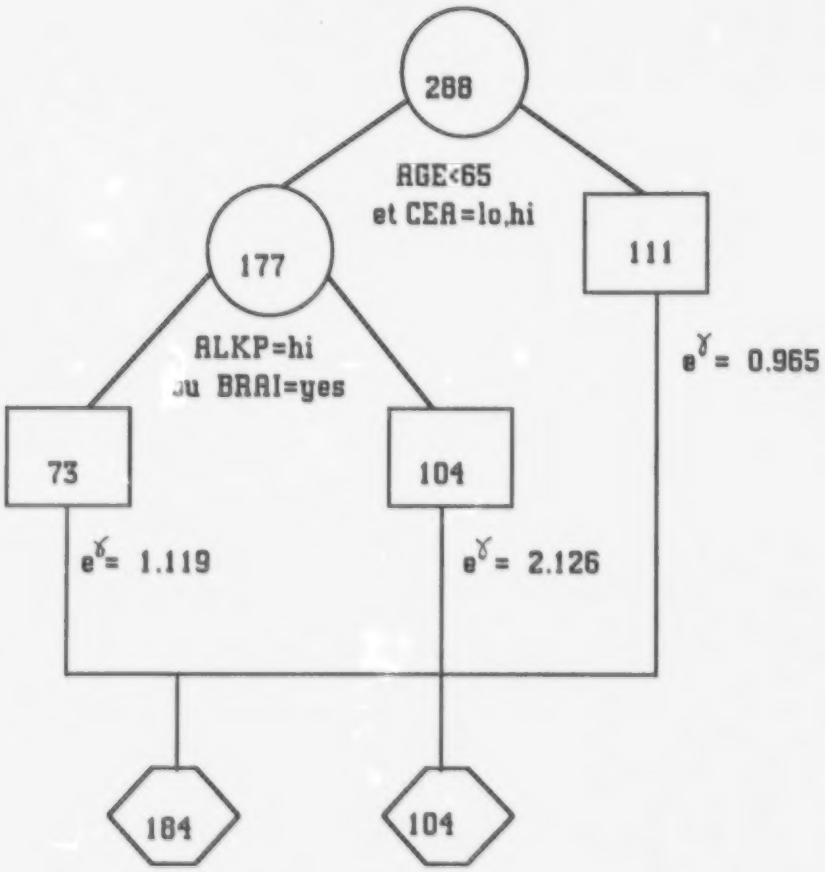

288
AGE<65
et CEA=lo,hi
177
111
ALKP=hi
ou BRAI=yes
e^{γ} = 0.965
73
104
e^{γ} = 1.119
e^{γ} = 2.126
184
104

Figure 5

STRATÉGIES D'INFÉRENCE STATISTIQUE EN RECHERCHE ÉPIDÉMIOLOGIQUE EN PRÉSENCE DE PARAMÈTRES DE NUISANCE

Jean Lambert
Université de Montréal

INTRODUCTION

Dans le cadre de recherches épidémiologiques, il est très fréquent de retrouver comme variables des indicateurs de morbidité ou de mortalité, donc des variables catégorielles. Nous nous en tiendrons donc au contexte de l'analyse de tableaux de contingence. Il faut cependant souligner que le problème de paramètres de nuisance n'est pas unique à l'analyse de tableaux de contingence. Par exemples: (1) une inférence sur μ lorsque σ est inconnu; (2) une inférence sur des coefficients de régression lorsque les variables indépendantes sont aléatoires; (3) une inférence sur deux moyennes lorsque les variances sont inégales; ... sont tous des exemples dans lesquels il y a présence d'un ou plusieurs paramètres de nuisance dans la distribution de la statistique d'intérêt.

Plusieurs solutions ont déjà été suggérées: (1) "Studentisation"; (2) transformation; (3) inférence fiduciaire; (4) inférence sur une sous-population; etc. (Hotelling, 1940; Cox, 1958; Rao, 1952; Kendall et Stuart, 1973).

Nous présentons plus loin les solutions suggérées dans un contexte d'analyses de tableaux de contingence après une brève introduction sur la problématique.

PROBLÉMATIQUE

Afin de simplifier la présentation, considérons un simple tableau de contingence (2 x 2). Comme première situation, l'on considère le test d'indépendance entre deux variables aléatoires caractérisant une distribution quadrinomiale dans un tableau (2 x 2). Nous utilisons la

notation courante avec les indices i et j et le signe "." dénotant une sommation sur cet indice.

Nous avons comme hypothèse d'indépendance:

$$H_0 : p_{ij} = p_{i\cdot}p_{\cdot j}$$

La probabilité d'observer un tel tableau de contingence sous H_0 est donnée par:

$$P\left(N_{ij}\middle|n_{\cdot\cdot},p_{1\cdot},p_{\cdot 1};H_0\right)=\binom{n_{\cdot\cdot}}{n_{ij}} p_{1\cdot}^{n_{1\cdot}}\,(1-p_{1\cdot})^{n_{2\cdot}}\,p_{\cdot 1}^{n_{\cdot 1}}\,(1-p_{\cdot 1})^{n_{\cdot 2}}\quad,\ (1)$$

Comme deuxième situation, l'on considère le test d'homogénéité entre deux probabilités binomiales.

Nous avons comme hypothèse d'homogénéité:

$$H_0 : p_{11} = p_{21} \quad (=p_\star)$$

La probabilité d'observer un tel tableau de contingence sous H_0 est donnée par:

$$P\left(N_{11},N_{21}\middle|n_{1\cdot},n_{2\cdot},p_\star;H_0\right) =\binom{n_{1\cdot}}{n_{11}}\binom{n_{2\cdot}}{n_{21}} p_\star^{n_{\cdot 1}}\,(1-p_\star)^{n_{\cdot 2}}\quad,\ (2)$$

On voit donc que les distributions d'échantillonnages (1) et (2) contiennent respectivement deux et un paramètre(s) de nuisance. Suivant l'approche de Neyman-Pearson, l'on observe que le test exact n'existe pas.

En présence de grands échantillons, l'on utilise le meilleur estimateur asymptotiquement normal (BAN estimator) par exemple, l'estimateur du maximum de vraisemblance (MLE estimator) qui a toutes les propriétés d'un

estimateur BAN, afin d'estimer le (ou les) paramètre(s) de nuisance. Il faut souligner qu'il existe une grande latitude si l'on s'intéresse plutôt à estimer la variance d'une statistique de test (e.g. la variance de $d = \hat{p}_{11} - \hat{p}_{21}$), puisque l'on peut assumer ou non l'hypothèse nulle comme étant vraie ou rechercher un estimateur non biaisé. Toutefois, toutes les procédures sont asymptotiquement équivalentes.

Cependant, en présence de petits échantillons, on peut questionner la qualité de tels estimateurs. Il est reconnu que, selon la situation sous-jacente, une statistique de test asymptotique peut mener à des décisions statistiques erronées (Lambert, 1987). Dans un tel contexte, nous présentons trois approches qui ont été proposées afin de solutionner ce problème:

1) Inférence conditionnelle.
2) "Elimination" des paramètres de nuisance.
3) Correction des procédures asymptotiques.

1. INFÉRENCE CONDITIONNELLE

Il existe un test uniformément plus puissant et non biaisé (UMPU test), tel que proposé par Tocher (1950), pour tester les hypothèses d'indépendance ou d'homogénéité dans un tableau de contingence (2 x 2). Ce test est un test conditionnel (test hypergéométrique ou test de Fisher), mais suppléé d'un processus de randomisation sur l'observation hypergéométrique adjacente à la région de rejet du test conditionnel, afin d'obtenir le seuil de signification nominal (α^*). Par exemple, dans la situation d'un test d'indépendance, le seuil de signification du test UMPU est donné par:

$$\alpha = \sum_{n_{1\cdot}=0}^{n_{\cdot\cdot}} \sum_{n_{\cdot 1}=0}^{n_{\cdot\cdot}} \alpha_{n_{1\cdot},n_{\cdot 1}} P(N_{1\cdot} = n_{1\cdot})\, P(N_{\cdot 1} = n_{\cdot 1}), \quad (3)$$

où, $\alpha_{n_{1\cdot}, n_{\cdot 1}}$ dénote la probabilité hypergéométrique de rejeter H_0, ayant observé $n_{1\cdot}$ et $n_{\cdot 1}$ et où les P ( ) dénotent les probabilités binomiales d'observer $n_{1\cdot}$ et $n_{\cdot 1}$ respectivement. Par un processus de randomisation, le seuil de signification du test conditionnel n'inclut pas seulement les observations hypergéométriques contenues dans la région de rejet, mais toute(s) observation(s) adjacente(s) avec une probabilité appropriée, de telle sorte que $\alpha_{n_{1\cdot}, n_{\cdot 1}} = \alpha^*$ pour chaque combinaison de $n_{1\cdot}$ et $n_{\cdot 1}$. Le seuil de signification du test est donc donné par:

$$\alpha = \alpha^* \sum_{n_{1\cdot}=0}^{n_{\cdot\cdot}} \sum_{n_{\cdot 1}=0}^{n_{\cdot\cdot}} P(N_{1\cdot} = n_{1\cdot})\, P(N_{\cdot 1} = n_{\cdot 1}) = \alpha^*$$

Cependant, même s'il s'agit d'un test UMPU, ce test n'est pas utilisé en pratique, étant donné qu'il repose sur un processus aléatoire complètement indépendant des données. Il a même déjà été qualifié d'approche "répugnante" pour l'analyse de données (Mantel, Greenhouse, 1968). Ce test représente cependant la référence lors d'investigations empiriques (e.g. simulations) de certains tests.

Donc, même en suivant la philosophie de Neyman-Pearson, la tendance a été d'utiliser l'équivalent non randomisé du test UMPU, conduisant à une procédure conservatrice (c.-à.-d. $\alpha < \alpha^*$) pour un seuil de signification nominal. Par exemple, si l'on utilise l'équivalent non randomisé pour un test unilatéral, avec $\alpha^* = 0.05$ et $n_{\cdot\cdot} = 6$ (très petit effectif pour simplifier les calculs), il existe un seul résultat significatif pour $n_{11} = 0$ (ou 3) et pour $n_{1\cdot} = n_{\cdot 1} = 3$. Donc, la formule (3) se réduit à:

$$\alpha = \alpha_{3,3}\, P(N_{1\cdot} = 3)\, P(N_{\cdot 1} = 3)$$

$$\alpha = 20\, p_{1\cdot}^{3} (1-p_{1\cdot})^{3}\, p_{\cdot 1}^{3} (1-p_{\cdot 1})^{3}$$

avec un $\alpha_{max} = 0.0049$ pour $p_{1\cdot} = p_{\cdot 1} = \frac{1}{2}$

Il s'agit donc d'une procédure très conservatrice. Quelques vérifications empiriques démontrent que:

a) Pour $n.. \leq 20$, on a $\alpha_{max} < \frac{1}{2} \alpha^*$

b) α_{max} diminue lorsque $p_{1.}$ (ou $p_{.1}$) tend vers 0 ou 1.

Des vérifications empiriques démontrent qu'il en est ainsi pour les autres devis de recherche fréquemment utilisés en épidémiologie ou recherche clinique:

- La comparaison de deux binomiales (échantillons indépendants).
- La comparaison de deux binomiales (mesures appariées).
- Le devis croisé (mesures appariées).
- La comparaison de deux variables Poisson.

L'inférence conditionnelle (e.g. le test de Fisher) n'est donc pas une solution au problème des paramètres de nuisance si l'on respecte l'approche de Neyman-Pearson.

2. "ELIMINATION" DES PARAMÈTRES DE NUISANCE

On peut subdiviser cette approche selon la méthode employée pour "éliminer" les paramètres de nuisance.

2.1 Estimation des paramètres de nuisance

Cette méthode fut proposée par Irwin (1935), mais étant donné les moyens de calcul de l'époque, elle fut mise de côté. Elle fut reprise, lorsque les moyens le permirent, par Salama (1974), Salama et Quade (1974) et Liddell (1976).

Considérons la comparaison de deux binomiales. Sous H_0, la distribution de probabilités de N_{11} et N_{21} est donnée par l'équation (2). En un premier temps, on estime le paramètre de nuisance (p_*) à l'aide d'un

estimateur du maximum de vraisemblance. En un deuxième temps, on choisit un critère d'ordre (C) approprié, qui reflète l'ensemble des hypothèses alternatives et qui possède certaines propriétés intéressantes (e.g. conditions de symétrie et de convexité; Barnard, 1947). Le chi-deux de Pearson, possédant ces propriétés, a été suggéré par Salama. En un troisième temps, on associe à ce critère d'ordre une distribution de probabilités empirique obtenue à l'aide de:

$$P\left(N_{11},N_{21}\middle|n_{1\cdot},n_{2\cdot},\hat{p}_{\star};H_0\right) = \binom{n_{1\cdot}}{n_{11}}\binom{n_{2\cdot}}{n_{21}}\hat{p}_{\star}^{\,n_{\cdot 1}}(1-\hat{p}_{\star})^{n_{\cdot 2}}$$

On rejette H_0 au seuil α^* si:

$$\sum_{C_i \geq C \text{ observé}} P\left(N_{11},N_{21}\middle|n_{1\cdot},n_{2\cdot},\hat{p}_{\star};H_0\right) \leq \alpha^*$$

Cette approche, tout en utilisant un estimateur pour le paramètre de nuisance et une statistique du chi-deux comme critère d'ordre, n'assume aucune fonction de répartition théorique, telle une distribution du chi-deux. Des vérifications empiriques faites par les auteurs démontrent que cette approche donne de meilleurs résultats que l'utilisation habituelle d'une statistique chi-deux ou d'un test conditionnel. Cependant, cette approche nécessite l'utilisation d'un logiciel ou de tables spécifiques.

2.2 Maximisation des paramètres de nuisance

On présente dans cette section le test de base (Barnard, 1947) sur lequel repose cette approche. Par la suite on présente le test de Boschloo (1970) qui constitue une amélioration du test de base.

Si l'on se réfère encore à la situation d'une comparaison de deux binomiales, Barnard (1947) suggéra de maximiser le seuil de signification rattaché à chaque tableau de contingence possible, et ce, en fonction du paramètre de nuisance p . Pour ce faire, il développa une notion d'ordre sur les tableaux de contingence, reflétant l'ensemble des hypothèses alternatives (i.e., conditions de symétrie et de convexité). Considérons la fonction de répartition suivante:

$$F\left(N_{11},N_{21}\middle|n_{1\cdot},n_{2\cdot},p_{\star};H_0\right) = \sum_{\text{extr.}} \binom{n_{1\cdot}}{n_{11}} \binom{n_{2\cdot}}{n_{21}} p_{\star}^{\,n_{\cdot 1}} (1-p_{\star})^{n_{\cdot 2}}$$

définie par la sommation des probabilités associées aux tableaux de contingence représentant des situations plus extrêmes que le tableau de contingence observé, et ce, en fonction de l'ensemble des hypothèses alternatives. Afin d'éliminer le paramètre de nuisance ($p_{\star}$), Barnard suggéra d'utiliser:

$$\alpha_{max} = \max_{0 < p_{\star} < 1} F\left(N_{11},N_{21}\middle|n_{1\cdot},n_{2\cdot},p_{\star};\ H_0\right)$$

Il est très intéressant d'observer que ce test, advenant un résultat significatif, est complètement informatif, en ce sens qu'aucune information externe sur le paramètre de nuisance ne peut mener à une autre conclusion statistique. Par contre, advenant un résultat non significatif, le test de Barnard est partiellement informatif : la connaissance de la valeur du paramètre de nuisance pourrait mener à une autre conclusion statistique.

Le test de Barnard est un test conservateur, en ce sens qu'il rejette H_0 si et seulement si cela est vrai pour toutes les valeurs de $p_{\star}$. Cependant, comme stipulé par Barnard, ce test est "progressif-conservateur" en comparaison du test hypergéométrique, puisqu'il donne la vraie borne supérieure de la probabilité de rejeter H_0.

Etant donné la nature discrète de la fonction de répartition, l'on peut concevoir aisément que α_{max} sera en général plus petit que α^*, menant à une perte de puissance du test. Ce phénomène étant plus important lorsque les effectifs d'échantillons sont petits. Afin de compenser cette perte de puissance, Boschloo (1970) suggéra d'élever α^* à un seuil α^1, de telle sorte que α_{max} se rapproche de α^* avec la restriction que $\alpha_{max} \leq \alpha^*$. En procédant de la sorte, des éléments sont rajoutés dans la région critique du test, menant à une augmentation de la puissance du test. Il faut noter qu'en procédant de la sorte, la puissance du test de Boschloo dépasse celle du test UMPU; la raison est que le test de Boschloo est un test biaisé. Boschloo publia des tables assez élaborées pour l'utilisation de son test.

McDonald, Davis et Milliken (1975, 1977) produisirent également de tables pour le test de Boschloo, qui ne nécessitent aucun calcul : (on présente les valeurs de n_{11} et n_{21} menant à des résultats significatifs pour différents effectifs et seuils de signification).

Comme pour la procédure de Salama (1974), le test de Boschloo nécessite donc l'utilisation d'un logiciel ou de tables spécifiques.

3. CORRECTION DES PROCÉDURES ASYMPTOTIQUES

Cette approche est une méthode indirecte, en ce sens qu'elle ne s'attaque pas directement au(x) paramètre(s) de nuisance. On essaie plutôt de corriger une statistique de test (e.g. un chi-deux), de telle sorte que son degré de signification (p) se rapproche le plus possible du degré de signification que l'on obtiendrait avec un test exact.

Il est reconnu que, dans un tableau de contingence (2 x 2), le chi-deux de Pearson constitue en général une procédure libérale ($\alpha > \alpha^*$) lorsque les effectifs d'échantillon sont petits (Camilli et Hopkins, 1978, 1979). Ce problème étant étroitement relié à l'approximation de probabilités d'une fonction de répartition discrète par une fonction

de répartition continue, il n'est pas surprenant de constater qu'une correction sur le chi-deux de Pearson à des fins d'une meilleure approximation d'un phénomène discret (e.g. probabilités quadrinomiales) rapprochera le seuil de signification empirique (α) du seuil de singification nominal (α^*).

Le concept de la correction de continuité fut développé par Yates (1934) qui améliora l'approximation normale d'une distribution hypergéométrique en introduisant un facteur de correction de $\frac{1}{2}$ qui représente la moitié de l'intervalle entre deux variables hypergéométriques. Dans un contexte d'analyse de tableaux de contingence il est important de souligner que la distribution hypergéométrique représente un tableau (2 x 2) avec tous les totaux marginaux fixes. Cette situation ne se présente jamais en recherche épidémiologique ou clinique où l'on s'intéresse plutôt à la comparaison de deux binomiales (ex.: les totaux colonnes seulement sont fixés) ou à l'analyse de l'indépendance de deux variables aléatoires caractérisant une distribution quadrinomiale (aucun total marginal fixé, sauf $n_{..}$).

Si l'on revient à la situation hypergéométrique, on peut décrire la méthode comme suit:

Soit S une variable hypergéométrique (<u>e.g.</u> n_{11}). $E(S|H_0)$ et $V(S|H_0)$ sont évidemment connues. L'intervalle entre deux S_i est 1. On obtient donc:

$$\chi^2_y = \frac{(|S - E(S|H_0)| - \frac{1}{2})^2}{V(S|H_0)}$$

$$\chi^2_y = \frac{(n_{..}-1)\,(|n_{11}n_{22}-n_{12}n_{21}|-n_{..}/2)^2}{n_{1.}n_{2.}n_{.1}n_{.2}}$$

Dans un contexte d'une analyse d'un tableau de contingence (2 x 2) caractérisé par une distribution hypergéométrique, ce chi-deux corrigé constitue en fait l'approximation asymptotique du test de Fisher (ou test hypergéométrique).

Dans les contextes produit-de-binomiales ou quadrinomiale, la situation est tout autre. Considérons par exemple un contexte de produit-de-binomiales. Nous avons comme statistique de test:

$$S = \frac{n_{11}}{n_{1\cdot}} - \frac{n_{21}}{n_{2\cdot}}$$

Sous l'hypothèse nulle, $E(S|H_0) = 0$.

Cependant,

$$V(S) = \frac{p_{11}(1-p_{11})}{n_{1\cdot}} + \frac{p_{21}(1-p_{21})}{n_{2\cdot}}$$

ou

$$V(S|H_0) = \frac{n_{\cdot\cdot}p_{\star}(1-p_{\star})}{n_{1\cdot}n_{2\cdot}}$$

On voit donc que $V(S)$ et $V(S|H_0)$ contiennent respectivement deux et un paramètre(s) de nuisance. Quel que soit notre choix entre $V(S)$ et $V(S|H_0)$ dans l'utilisation d'un chi-deux, nous aurons la latitude entre un estimateur biaisé (MLE) ou non biaisé.

Par exemple, l'utilisation d'un estimateur non biaisé pour $V(S|H_0)$, c'est-à-dire :

$$\tilde{V}(S|H_0) = \frac{n_{..}^2 \hat{p}_{.1}(1-\hat{p}_{.1})}{(n_{..}-1)n_{1.}n_{2.}}$$

nous conduit à la statistique suivante :

$$\chi^2 \simeq \frac{(n_{..}-1)\ (n_{11}n_{22} - n_{12}n_{21})^2}{n_{1.}n_{2.}n_{.1}n_{.2}}$$

Il est important de souligner que l'intervalle entre deux S_i successives n'est pas constant. Egalement, quoique la distribution de probabilités conjointe de n_{11} et n_{21} soit un produit-de-binomiales, la distribution de probabilités de S n'est pas binomiale. L'étude empirique de la distribution de S nous permet cependant de déterminer l'intervalle moyen entre deux S_i successives. Si l'on définit par k le plus grand diviseur commun de $n_{1.}$ et $n_{2.}$, cet intervalle moyen est donné par:

$$I = \frac{k(2n_{1.}n_{2.}+k)}{n_{1.}n_{2.}(2k-1) + kn_{..}}$$

L'intégration de la moitié de cet intervalle dans un chi-deux corrigé conduirait cependant à une expression un peu lourde, dans laquelle la correction de continuité (cc) serait donnée par:

$$cc = \frac{k^2(n_{1.}n_{2.}+k/2)}{n_{1.}n_{2.}(2k-1) + kn_{..}}$$

Une approximation conservatrice (sous-estimation) de l'intervalle moyen nous conduit cependant à une formule très simple pour un chi-deux corrigé :

$$\chi^2_c = \frac{(n..-1)\,(|n_{11}n_{22} - n_{12}n_{21}| - k/2)^2}{n_1.n_2.n._1n._2}$$

Nous avons utilisé ici un estimateur non biaisé pour $V(S|H_0)$, afin de faire le parallèle avec χ^2_y.

Il est intéressant de noter que si $n_2.$ est un entier multiple de $n_1.$, la cc se réduit à :

$$\frac{k^2(n_1.n_2.+k/2)}{n_1.n_2.(2k-1) + kn..} = \frac{k}{2} = \frac{n_1.}{2}$$

Egalement, si $n_1. = n_2.$,

$$\frac{k^2(n_1.n_2.+k/2)}{n_1.n_2.(2k-1)+kn..} = \frac{k}{2} = \frac{n..}{4}$$

Dans le cas où $n_1. = n_2.$, une correction de continuité de n../4 fut présentée par Pirie et Hamdan (1972), qui étudièrent également d'autres situations.

Il est important de souligner qu'une cc doit être fonction de la distribution de probabilités sous-jacente. Si l'on compare les cc pour les distributions hypergéométrique et produit-de-binomiales, on remarque que la cc pour la distribution produit-de-binomiales sera

toujours inférieure à la cc pour la distribution hypergéométrique : ce qui ne doit pas nous surprendre, puisque la distribution produit-de-binomiales est une distribution beaucoup moins discrète que l'hypergéométrique.

Des simulations (5 000 répétitions) pour chacune des conditions suivantes:

a) $2(1)9 \leq n_{1.} \leq n_{2.} \leq 20$
b) $p_* = 0.1, 0.2, 0.3, 0.4, 0.5$
c) $\alpha^* = 0.01, 0.05$

ont démontré que X_c^2 doit être préféré au chi-deux non corrigé qui est trop libéral. Quand au test de Fisher ou à son approximation asymptotique, c'est-à-dire X_y^2 , ils présentent des résultats inacceptables, étant donné leur fort conservatisme, d'où une perte de puissance importante.

CONCLUSION

Nous pouvons considérer que, en présence de petits échantillons dans un tableau de contingence, la correction d'une procédure asymptotique représente la solution la plus simple et la plus accessible au problème des paramètres de nuisance.

RÉFÉRENCES

Barnard, G.A. (1947). "Significance Tests for 2x2 Tables". Biometrika, 34, 123-138.

Boschloo, R.D. (1970). "Raised Conditional Level of Significance for the 2x2 Table when Testing the Equality of two Probabilities". Statistica Neerlandica, 24, 1-35.

Camilli, G. and Hopkins, K.D. (1978). "Applicability of Chi-Square to 2x2 Contingency Tables with Small Expected cell Frequencies". Psych. Bull., 85, 163-167.

Camilli, G. and Hopkins, K.D. (1979). "Testing for Association in 2x2 Contingency Tables with Very Small Sample Sizes". Psych.Bull. 86, 1011-1014.

Cox,D.R. (1958). "Some Problems Connected with Statistical Inference". Ann. Math. Statist., 29, 357-372.

Hotelling, H. (1940). "The Selection of Variates for Use in Prediction with Some Comments on the General Problem of Nuisance Parameters". Ann. Math. Statist., 11, 271-283.

Irwin, J.O. (1935). "Tests of Significance for Differences between Percentages Based on Small Numbers". Metron, 12, 83-94.

Kendall, M.G. and Stuart, A. (1973). The Advanced Theory of Statistics, 3rd ed., Hafner Press, New York.

Lambert, J. (1987). "Quelques réflexions sur l'inférence statistique dans l'analyse de données catégorielles". Actes du Colloque sur les Applications de la Statistique, Bureau de la statistique du Québec, pp. 251-260.

Liddell, D. (1976). "Practical Tests of 2x2 Contingency Tables". The Statistician, 25, 295-304.

Mantel, N. and Greenhouse, S.W. (1968). "What is the Continuity Correction?" The Am. Statistician, 22, 27-30.

McDonald, L.L., Davis, B.M. and Milliken, G.A. (1975). "A Nonrandomized Unconditional Test for Comparing Two Proportions". Faculty Research Paper No. 94, Stat. Lab. #543, Dept. of Statistics, University of Wyoming.

McDonald, L.L., Davis, B.M. and Milliken, G.A. (1977). "A Nonrandomized Unconditional Test for Comparing Two Proportions in 2x2 Contingency Tables". Technometrics, 19, 145-157.

Pirie, W.R. and Hamdan, M.A. (1972). "Some Revised Continuity Corrrections for Discrete Distributions". Biometrics, 28, 693-701.

Rao, C.R. (1952). Advanced Statistical Methods in Biometric Research, Hafner Press, New York.

Salama, I.A. (1974). "Contributions to the Theory of Statistical Inference from Two-Dimensional Contingency Tables". Department of Biostatistics, University of North Carolina at Chapel Hill, Institute of Statistics Mimeo Series No. 924. (1-155).

Salama, I.A. and Quade, D. (1974). "Significance Levels of a New Test for Two-By-Two Tables". Department of Biostatistics, University of North Carolina at Chapel Hill, Institute of Statistics Mimeo Series No. 937. (1-175).

Tocher, K.D. (1950). "Extension of the Neyman-Pearson's Theory of Tests to Discontinuous Variates". Biometrika, 37, 130-144.

Yates, F. (1934). "Contingency Tables Involving Small Numbers and the X^2 Test". J. R. Stat. Soc.., Suppl. 1, 217-235.

SESSION SUR LA STATISTIQUE NON PARAMÉTRIQUE

Estimation et Intervalle de Confiance Adaptable Pour un Paramètre de Localisation

Luc Perreault
Université Laval

Lorsque la loi d'où proviennent les observations est inconnue les méthodes d'estimation et de construction d'intervalles de confiance classiques pour un paramètre de localisation risquent de donner de moins bons résultats. Pour pallier ce problème, d'autres méthodes ont été développées. Dans ce travail, il est question de construction d'intervalles de confiance adaptables. L'objectif principal de cette méthode est d'adapter la forme de l'intervalle à la nature des données. Nous présentons et comparons deux intervalles de confiance adaptables. Enfin, nous proposons un intervalle complètement valide et qui possède une bonne efficacité.

1. INTRODUCTION

Dans ce travail, nous nous intéressons à la construction d'intervalle de confiance pour un paramètre de localisation lorsque la taille d'échantillon est petite. L'objectif principal est d'obtenir un intervalle de confiance qui soit bon pour des critères de qualité bien précis. Deux critères sont traditionnellement introduits: la validité et l'efficacité. Un intervalle de confiance est dit valide si son niveau de confiance réel est égal au niveau de confiance nominal (celui que l'on s'est fixé avant l'expérience). Il est dit efficace s'il a une grande précision, c'est-à-dire qu'il n'est pas trop long.

Considérons le modèle nonparamétrique de localisation que l'on définit de la façon suivante.

Soit $\mathbf{F}_s$ *l'ensemble des lois absolument continues et symétriques par rapport à l'origine. Il existe* $F \in \mathbf{F}_s$ *et* $\theta \in \Theta \in \mathbb{R}$ *tel que* $X_1, \ldots, X_n$ *est un échantillon de la loi* F_θ *définie par*

$$\forall x \in \mathbb{R}, \quad F_\theta(x) = F(x - \theta). \tag{1.1}$$

L'hypothèse de continuité est faite que pour faciliter les développements mathématiques. L'hypothèse de symétrie permet d'interpréter de façon simple et naturelle le paramètre de localisation. Il représente alors soit la moyenne, si

elle existe, soit la médiane, ou encore toute caractéristique de tendance centrale de la loi comme, par exemple, les moyennes α-tronquées. On emploie ce type de modèle lorsqu'on ne connait pas exactement la loi d'où proviennent les observations. Face à une telle situation, on peut aborder le problème de la construction d'intervalles de confiance de deux façons.

La première approche consiste à choisir l'estimateur et l'intervalle de confiance sans regarder les observations. Par exemple, on peut décider avant la cueillette des données d'estimer θ par la moyenne $\bar{X}_n$ et d'utiliser un intervalle de confiance de Student. Evidemment, nous espérons que les observations proviennent de la loi normale puisque dans ce cas nous aurons choisi la solution optimale. Sachant que l'intervalle de confiance de Student peut être mauvais pour une loi autre que la loi normale, on pourrait, toujours sans regarder notre échantillon, utiliser un intervalle de confiance non paramétrique ou robuste optant ainsi pour la prudence.

La deuxième façon d'aborder le problème est d'utiliser les observations pour guider notre choix d'un estimateur et d'un intervalle de confiance. Ce type d'approche est connu sous le nom de procédure adaptable. Beaucoup de praticiens de la statistique emploi ce genre de méthode. Souvent, par exemple, on choisit d'utiliser un intervalle de confiance de Student plutôt qu'un intervalle de confiance non paramétrique après avoir testé la normalité. Il existe plusieurs exemples semblables à celui-ci. C'est pourquoi, on dit souvent que la théorie des méthodes adaptables est la formalisation de ce que beaucoup de statisticiens font en pratique.

Plus généralement, une procédure adaptable consiste à utiliser une statistique, que l'on appelle adaptative, nous renseignant sur la loi d'où proviennent les observations pour choisir une bonne solution (estimateur, intervalle de confiance, test, etc...). On se sert donc deux fois des observations, une première fois pour essayer d'identifier la loi à l'aide de la statistique adaptative, et une seconde fois pour appliquer la solution choisie.

Pour construire un intervalle de confiance adaptable, on peut considérer un certain nombre d'intervalles d'une même famille. Parmi les intervalles retenus, on choisit le meilleur pour la loi identifiée à l'aide de la statistique adaptative.

Dans ce qui suit, nous présentons premièrement deux statistiques adaptatives, l'une proposée par Hogg (1974) et l'autre par Policello et Hettmansperger (1976). Nous décrivons ensuite une famille d'intervalles de confiance

nonparamétriques parmi lesquels on en sélectionne trois. Nous obtenons ainsi deux intervalles de confiance adaptables que nous comparons au niveau de l'efficacité et de la validité. On s'aperçoit que ces deux intervalles ne sont pas valides. Nous terminons donc en proposant une façon d'obtenir un intervalle de confiance adaptable complètement valide.

2. STATISTIQUES ADAPTATIVES

Pour identifier la loi d'où proviennent les observations, nous avons besoin d'une caractéristique qui différencie les lois dans $\mathbf{F}_r$. Nous utilisons comme caractéristique le poids aux extrémités des lois. Cette caractéristique de la forme des lois est reconnue depuis longtemps comme ayant une grande influence sur certains estimateurs et intervalles de confiance classiques.

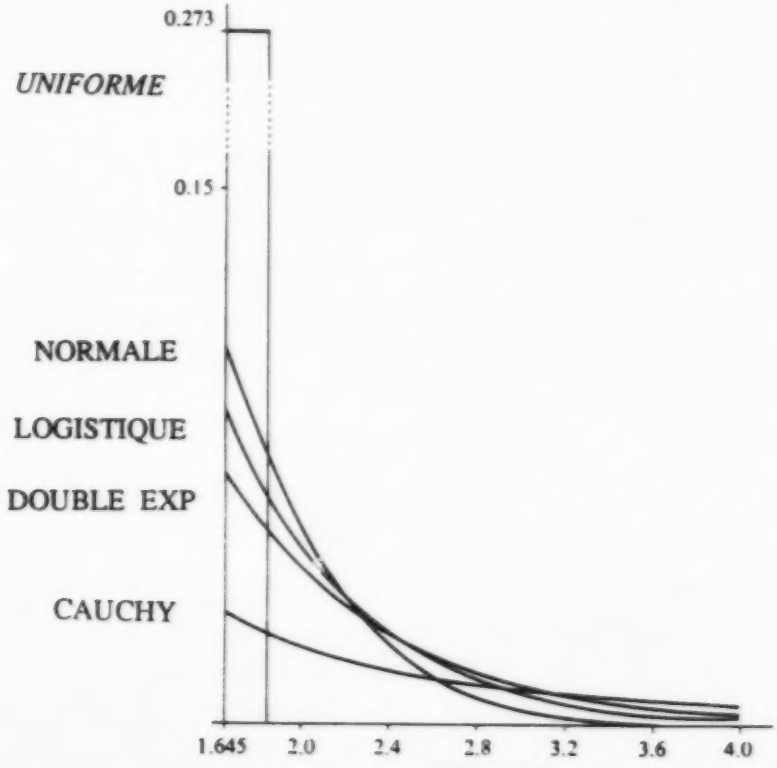


Figure 2.1

De façon intuitive, on considère qu'une loi a des extrémités lourdes si elle contient relativement plus de probabilité dans ses extrémités que dans ses

régions centrales. La figure 2.1 montre comment peuvent se placer l'une par rapport aux autres les densités de quelques lois dont les poids aux extrémités diffèrent. Ce graphique représente les extrémités droites de cinq lois que l'on a standardisées de façon à ce qu'elles aient le même $95^{ième}$ percentile. On remarque, par exemple, que la probabilité d'obtenir une valeur extrême de 4 est presque égale à celle d'obtenir une valeur plus centrale de 1.645 pour la loi de Cauchy. A l'inverse, pour la loi uniforme, la probabilité d'observer une valeur plus grande que 1.8 est nulle tandis que celle d'observer 1.645 est 0.273. On dit dans ce cas que la loi de Cauchy a des extrémités plus lourdes que la loi uniforme. On peut, à l'aide de cette caractéristique, ordonner quelques lois. Par exemple,

UNIFORME <. NORMALE <. LOGISTIQUE <. DOUBLE EXPONENTIELLE

où <. signifie "a des extrémités plus légères que".

La statistique adaptative qui nous permet d'identifier à peu près la loi et de choisir un intervalle de confiance doit mesurer cette caractéristique. R.V. Hogg (1974) propose la statistique

$$Q = \frac{\overline{U}(0.2) - \overline{L}(0.2)}{\overline{U}(0.5) - \overline{L}(0.5)} , \qquad (2.1)$$

où $\overline{U}(\beta)$ et $\overline{L}(\beta)$ sont respectivement les moyennes des $n\beta$ plus grandes et $n\beta$ plus petites statistiques d'ordre. Q nous donne le poids relatif des observations extrèmes de l'échantillon par rapport au poids des observations centrales. Cette statistique mesure donc la caractéristique qui nous intéresse.

Policello et Hettmansperger (1976) suggère plutôt l'utilisation d'une statistique adaptative basée directement sur les intervalles de confiances sélectionnés. Ils proposent tout simplement de prendre l'intervalle de confiance le plus court. Plus précisément, leur statistique adaptative est

$$L_i = \Delta^i_{(U)} - \Delta^i_{(L)}{}^{*} \qquad i = 1, ..., p \qquad (2.2)$$

où $\Delta^i_{(L)}$ et $\Delta^i_{(U)}$ sont respectivement la borne inférieure et supérieure de l'intervalle i, et p le nombre d'intervalles de confiance retenus dans la famille choisie.

---

* L_i est à une constante multiplicative près un estimateur de la variance asymptotique de la statistique linéaire de rangs du test à partir duquel est construit l'intervalle i (voir Sen, 1966).

3. INTERVALLES DE CONFIANCE DE GARSWIRTH

Une fois la statistique adaptative bien déterminée, il suffit de sélectionner un certain nombre d'intervalles de confiance. Pour ce faire, nous proposons d'utiliser une famille d'intervalles nonparamétriques introduite par Garswirth (1965). Ces intervalles de confiance sont déduits de tests de rangs signés. Ils ont donc l'avantage d'avoir un niveau de confiance constant pour un vaste ensemble de lois contrairement à ceux construits par la méthode du pivot.

Une façon simple de calculer les bornes de tels intervalles est de former le tableau suivant où les marges contiennent les n observations en ordre croissant et où, à l'intérieur, sont mises les valeurs des demi-sommes $\Delta_k = (X_i + X_j)/2$.

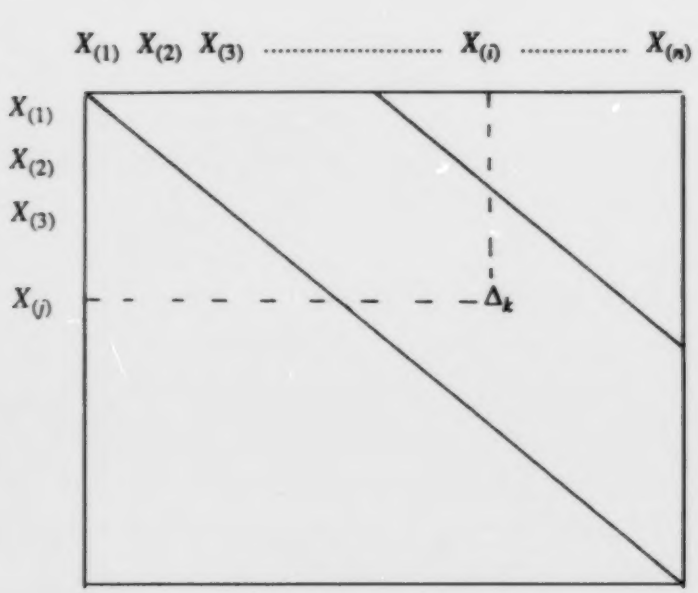


Table 3.1

On ne considère ensuite que les $n(1-\nu_1)$ plus grandes observations ce qui correspond à des demi-sommes situées dans un triangle supérieur du tableau, ou que les $n(1-\nu_2)$ plus petites ce qui correspond à des demi-sommes qui se retrouvent dans un trapèze inférieur du tableau. Les bornes de l'intervalle sont alors calculées à l'aide d'un de ces deux sous-ensembles de demi-sommes que l'on a ordonnées. L'intervalle de confiance résultant s'écrit

$$I_{\nu_1,\nu_2} = [\ \Delta_{(k)}\ ,\ \Delta_{(k')}\ ] \qquad 0 \le \nu_1,\ \nu_2 \le 1 \tag{3.1}$$

tel que

$$P\{\theta \in\ [\ \Delta_{(k)}\ ,\ \Delta_{(k')}\ ]\} = 1 - \alpha$$

où $\Delta_{(k)}$ est la $k^{ième}$ plus grande demi-somme non tronquée et $1 - \alpha$ le niveau de confiance nominal.

Cette famille d'intervalles indicés par ν_1 et ν_2 possède une propriété intéressante. On peut faire varier la proportion de troncature ($n\nu_1$ ou $n\nu_2$) pour obtenir un bon intervalle de confiance tout en suivant le poids aux extrémités des lois. En effet, si l'on enlève des demi-sommes situées dans un trapèze inférieur du tableau ($\nu_1 > 0$, $\nu_2 = 0$), l'intervalle obtenu est bon pour une loi aux extrémités plus légères que celles de la loi logistique. A l'inverse, si l'on enlève des demi-sommes contenues dans un triangle supérieur ($\nu_2 > 0$, $\nu_1 = 0$), l'intervalle est bon pour une loi aux extrémités plus lourdes que celles de la loi logistique. Enfin, lorsque ν_1=ν_2=0, l'intervalle de confiance résultant est celui déduit du test de Wilcoxon optimal pour cette même loi.

Nous sélectionnons dans cette famille trois intervalles de confiance, I_1 ($\nu_1 = 0.5$, $\nu_2 = 0$) qui sera bon pour des lois aux extrémités légères, I_2 ($\nu_1 = 0$, $\nu_2 = 0$) déduit du test de Wilcoxon et I_3 ($\nu_2 = 1+1/n$, $\nu_1 = 0$) l'intervalle déduit du test du signe, optimal pour la loi double exponentielle. Les deux intervalles de confiance adaptables que nous proposons sont alors les suivants,

$$IC_{HO} = \begin{cases} I_1 & si\ Q < 1.70^{*}, \\ I_2 & si\ 1.70 \le Q \le 1.87, \\ I_3 & si\ Q > 1.87 \end{cases} \tag{3.2}$$

et

$$IC_{PO} = I_i \tag{3.3}$$

où i est tel que $L_i = \min\{\ L_1, L_2, L_3\ \}$.

---

* Q prend théoriquement les valeurs 1.75, 1.81 et 1.93 respectivement pour les lois Normale, Logistique et Double Exponentielle.

4. ETUDE COMPARATIVE

Dans cette section, on compare les intervalles de confiance adaptables au niveau de l'efficacité et de la validité. IC_{HO} et IC_{PO} ne diffèrent que par la statistique adaptative. C'est pourquoi cette étude comparative porte sur les statistiques Q et L_i.

4.1. Efficacité

Du point de vue de l'efficacité, la statistique adaptative L_i semble préférable à Q puisqu'elle nous assure du choix de l'intervalle le plus court. Pour mieux étudier ce critère, nous avons procédé à une étude empirique. Nous avons généré 200 échantillons aléatoires de tailles n = 10, 20 et 40 provenant de quatre lois différentes ordonnées selon leur poids aux extrémités. Nous avons ensuite compilé le nombre de fois que chacun des intervalles I_1, I_2 et I_3 furent sélectionnés par Q et L_i. Cette simulation nous permet de voir si les deux statistiques sélectionnnent l'intervalle approprié pour chacune des lois. Nous vous présentons dans la table qui suit les résultats odtenus pour n = 10.

L_i, Q	UNIF.	NORM.	LOGIS.	D. EXP.
I_1	152, 144	97, 88	68, 66	55, 48
I_2	29, 45	51, 65	77, 81	59, 68
I_3	19, 11	52, 47	55, 53	86, 84

Table 4.1

Malgré la petite taille d'échantillon, les deux statistiques sélectionnent majoritairement le meilleur intervalle dans chacune des situations. Par exemple, I_1, bon pour des lois aux extrémités légères, est choisi 152 fois par L_i et 144 fois par Q lorsque les observations proviennent d'une loi uniforme. I_2 et I_3 sont aussi sélectionnés majoritairement par les deux statistiques lorsque la loi est celle pour laquelle ces intervalles sont optimaux. On remarque, et ce pour les trois tailles d'échantillon, que la statistique L_i semble choisir un peu plus souvent l'intervalle approprié que Q pour des lois extrêmes comme la loi uniforme et la loi double exponentielle.

4.2. Validité

Les deux intervalles de confiance IC_{HO} et IC_{PO} ne sont pas en général valides. Ceci est dû au fait que l'on utilise les observations deux fois de façon non indépendante. En effet, elles sont utilisées une première fois pour calculer Q ou L_i et une seconde fois pour calculer les bornes de l'intervalle choisi. Les statistiques adaptatives n'étant pas en général indépendantes de l'intervalle sélectionné, le niveau de confiance réel est différent de celui que l'on se fixe avant l'expérience. Il est en fait conditionnel à la réalisation de la statistique adaptative.

En résumé, pour qu'un intervalle de confiance adaptable soit bon pour les critères qui nous intéressent, la statistique adaptable doit posséder deux qualités. Elle doit bien identifier la loi d'où proviennent les observations de façon à choisir l'intervalle le plus efficace, et doit dépendre le moins possible de celui-ci afin que l'intervalle soit valide.

5. SOLUTION PROPOSEE

Pour obtenir un intervalle de confiance adaptable valide, une solution naturelle serait de tirer aléatoirement un sous-échantillon de taille $n/2$ et de calculer la statistique adaptative à l'aide de celui-ci. On utiliserait le reste de l'échantillon pour évaluer les bornes de l'intervalle sélectionné. De cette façon, la statistique adaptative serait indépendante de l'intervalle choisi et la solution serait valide. Il y aurait par contre une perte d'efficacité pour les échantillons de petite taille. En effet, moins il y a d'observations plus la statistique adaptative a de la difficulté à bien identifier la loi.

Pour éviter une telle perte d'efficacité, nous proposons une autre solution. Celle-ci s'appuie sur le fait que le vecteur de rangs signés $(R_1^+, R_2^+, ..., R_n^+)$ est indépendant du vecteur des observations en valeurs absolues ordonnées $(|X|_{(1)}, |X|_{(2)}, ..., |X|_{(n)})$ (voir Capéràa et Van Cutsem, 1988). Les intervalles de confiance que nous utilisons étant déduit de tests de rangs signés, ils sont indépendants de toute fonction des observations en valeurs absolues. Nous proposons donc de calculer la statistique adaptative L_i (elle semble en générale aussi bonne sinon meilleur que celle de Hogg) à l'aide des observations transformées et les bornes de l'intervalle choisi avec l'échantillon original. De cette façon, l'intervalle de confiance adaptable est valide et l'identification de la loi se fait à l'aide de tout l'échantillon au lieu d'un sous-ensemble.

Lorsque $\theta = 0$, la loi des observations transformées en valeurs absolues G est la même que celle des observations originales F. Par contre, plus θ augmente ou diminu plus G s'éloigne de la loi qui nous intéresse F. Si notre paramètre de localisation est très élevé nous aurons donc une perte d'efficacité.

6. CONCLUSION

Les deux solutions proposées nous donnent des intervalles de confiance valides. Par contre, il y a perte d'efficacité. La procédure qui consiste à transformer les observations en valeurs absolues semble plus efficace pour n=10 et celle qui consiste à diviser l'échantillon en deux semble meilleure pour n=20 et n=40. De plus, des simulations démontrent qu'il est préférable d'utiliser l'une de ces deux approches adaptables plutôt que de faire le choix d'un intervalle sans regarder les observations. Enfin, nous croyons qu'il est possible d'améliorer la statistique adaptative L_i et ainsi obtenir des intervalles de confiance adaptables plus efficaces. Des études empiriques sont en cours.

REFERENCES

[1] Garswirth, J.L. (1965). Percentile Modified Rank Tests. *J. Amer. Statist. Assoc.*, 60, 1127-41.

[2] Capéràa, Ph. et Van Cutsem, B. (1988). *Méthodes et modèles en statistique non paramétrique: Exposé fondamental.* Dunod, Paris.

[3] Hogg, R.V. (1974). Adaptive robust procedures: A partial review and some suggestions for future applications and theory. *J. Amer. Statist. Assoc.*, 69, 909-927.

[4] Policello, G.E.et Hettmansperger, T.P. (1976). Adaptative Robust Procedure for the One-Sample Location Problem. *J. Amer. Statist. Assoc.*, 71, 624-633.

[5] Sen, P.K. (1966). On a Distribution-Free Method of Estimating Asymptotic Efficiency of a Class of Non-Parametric Tests. *Ann. Math. Statist.*, 37, 1759-70.